의학박사 학위논문

안저영상 기반 녹내장 진단 및 중증도 등급화를 위한 딥러닝 모델

Deep learning model for glaucoma diagnosis and its stages classification based on fundus images

2019년 2월

서울대학교 대학원

의학과 의료정보학 전공

조 현 성

안저영상 기반 녹내장 진단 및 중증도 등급화를 위한 딥러닝 모델

지도교수 김 홍 기

이 논문을 의학박사 학위논문으로 제출함

2018년 10월

서울대학교 대학원
의학과 의로정보학 전공
조 현 성

조현성의 박사 학위논문을 인준함

2019년 1월

위 원 장 김 석 화 (인)

부위원장 김 홍 기 (인)

위 원 김 정 은 (인)

위 원 박 기 호 (인)

위 원 정 교 일 (인)

초록

서론: 본 연구는 안저영상을 바탕으로 녹내장 선별검사와 녹내장 중증도를 자동으로 분류하기 위한 합성곱신경망의 앙상블 방법에 관한 것이다. 녹내장의 선별검사와 중증도 등급화를 자동화하기 위해 서로 다른 특성을 갖는 48개의 합성곱신경망 모델을 정의하고 훈련했다. 학습을 완료한 모든 모델은 본 연구에서 제안하는 앙상블 방법을 통해서 최종 판독 결과를 도출하였고, 그 성능을 평가하였다.

방법: 본 연구에서는 합성곱신경망 모델의 훈련을 위해 2,801명의 환자로부터 측정한 4,445장의 안저영상을 수집하였다. 수집한 안저영상은 4명의 녹내장 전문의가 정상 집단과 녹내장 집단으로 분류하고, 녹내장 집단은 시야검사 결과의 평균 편차 (Mean Deviation, MD)를 참조하여 초기 녹내장 집단과 중증 녹내장 집단으로 세분화하였다. 이때, 평균 편차가 -6dB 이하를 중증 녹내장 집단으로 분류하였다. 또한, 환자 1명으로부터 좌, 우 각각 최대 1장씩의 안저영상을 사용하였다. 전체 안저영상 중에서 영상 품질이 열악한 것과 4명 녹내장 전문의의 등급 판정 결과가 100% 일치하지 않은 영상을 제외한 2,204명의 3,460장을 가지고 합성곱신경망 모델을 훈련하였다.
모델의 성능은 정확도, 민감도, 특이도, AUROC(Area Under the Receiver Operating Characteristic)을 평가 지표로 삼았다. 이때, InceptionNet-v3를 기준 모델로 하고, 본 연구에서 제안한 앙상블 방법과 성능을 비교하였다. 두 방법의 성능평가 결과를 Shapiro-Wilk normality test로 정규성 검정을 하였으며, paired t-test를 사용하여 두 방법의 성능 차이에 대한 통계적 유의성을 검정하였다.

결과: 본 연구에서 제안한 합성곱신경망 앙상블 방법은 녹내장 선별검사에 관한 것과 녹내장 중증도 분류에 관한 것으로 분리하여 성능을 평가하였다. 녹내장 선별검사의 정확도 측면에서 앙상블 방법은 96.6% (95% confidence interval [CI], 95.5 ~ 97.8%)를 보였다. 반면, InceptionNet-v3

모델 한 개를 사용한 기준 모델은 93.9% (95% CI, 92.6 ~ 95.2%)를 보였다. 기준 모델과 앙상블 방법의 녹내장 선별검사 정확도에 대한 성능 차이는 paired t-test를 통해 통계적 유의성을 검정하였고, 그 결과는 p-value 0.000425로 정확도의 차이가 통계적으로 유의함을 밝혔다. AUROC 측면에서 앙상블 방법은 0.994 (95% CI, 0.990 ~ 0.997)를 보였으며, InceptionNet-v3 모델 한 개를 사용한 기준 모델은 0.977 (95% CI, 0.969 ~ 0.986)를 보였다. 녹내장 선별검사에 있어서 기준 모델과 앙상블 방법의 AUROC에 대한 성능 차이는 역시 paired t-test를 통한 통계적 유의성을 검정하였고, 결과는 p-value 0.000966으로 AUROC의 차이가 통계적으로 유의함을 밝혔다. 이로써 녹내장 선별검사에서 앙상블 방법이 정확도와 AUROC 측면에서 더 높고 안정적인 것을 확인하였다.

녹내장 중증도 분류의 정확도 측면에서 앙상블 방법은 87.7% (95% CI, 85.9 ~ 89.7%)를 보였고, InceptionNet-v3 모델 한 개를 사용한 기준 모델은 82.3% (95% CI, 80.2 ~ 84.1%)를 보였다. 녹내장 중증도 분류에 있어서 기준 모델과 앙상블 방법의 정확도 차이는 paired t-test를 통해 통계적 유의성을 검정하였고, 그 결과는 p-value 0.002902로 그 차이가 통계적으로 유의함을 밝혔다. 평균 AUROC 측면에서 앙상블 방법은 0.975 (95% CI, 0.967 ~ 0.983)를 보였으며, InceptionNet-v3 모델 한 개를 사용한 기준 모델은 0.938 (95% CI, 0.926 ~ 0.949)을 보였다. 녹내장 중증도 분류에 있어서 평균 AUROC에 대한 기준 모델과 앙상블 방법의 성능 차이 역시 paired t-test를 통해 통계적 유의성을 검정하였고, 그 결과는 p-value 0.000093으로 그 차이가 통계적으로 유의함을 밝혔다. 이로써 녹내장 중증도 분류에서도 앙상블 방법이 정확도와 AUROC 측면에서 더 높고 안정적인 것을 확인하였다.

결론: 본 연구에서 제안하는 여러 개의 합성곱신경망을 앙상블 하는 방법은 안저영상을 바탕으로 녹내장 선별검사와 중증도 분류를 자동화하는데 있어서 기존의 방법보다 우수하고 안정적인 성능을 발휘한다. 본 연구결과는 인공지능 기술을 바탕으로 하는 임상 의사 결정 지원 시스템 (Clinical Decision Support System, CDSS) 소프트웨어로, 현재 널리 보급된 안저촬영기에 탑재 또는 연동하는 방식으로 다양한 분야에서 활용할 수 있다. 안저촬영기에 본 연구결과를 탑재하여 건강검진센터나 안과 진

료현장에서 활용한다면, 안저촬영 결과의 판독 효율과 정확성을 높일 수 있고, 이에 따른 시간적 이득을 전문의의 2차 판독에 할애함으로써 보다 경제적이고 정확한 검진 결과를 얻을 수 있다. 또한, 본 연구결과를 활용한 의료서비스가 활성화된다면, 잠재적 녹내장 환자에 대한 조기진단의 가능성을 높일 수 있고, 이에 따른 녹내장 환자의 치료 효과 향상과 관련 의료비용 지출을 절감할 수 있다.

주요어 : 녹내장, 안저영상, 딥러닝, 합성곱신경망, 앙상블, 인공지능

학 번 : 2011-30639

목 차

표 목 차

그 림 목 차

제 1 장 서론

제 1 절 연구의 배경과 의의

녹내장(Glaucoma)은 신경절세포(Retinal Ganglion Cell, RGC)와 그 축삭(Axon)이 손상됨으로써 실명에 이를 수 있는 질병이다 (Tham et al., 2014). 녹내장은 만성적이고, 비가역적으로 진행하는 특성 때문에 조기에 발견한 녹내장은 치료제 또는 수술을 통해 진행을 늦출 수 있고, 치료 효과 역시 좋은 편이다. 하지만, 질환의 특성상 말기 단계까지 환자가 뚜렷하게 느낄 수 있는 시야 결손이나 시력 저하와 같은 주관적인 증상이 없는 경우가 많다. 또한, 진행 단계가 심한 녹내장의 경우 치료 예후가 좋지 않아서, 검진을 통한 조기 발견의 중요성이 매우 크다.

최근 인구 조사에 근거한 50건의 글로벌 메타 분석에 따르면, 전 세계적으로 40세부터 80세까지의 녹내장 유병률은 약 6천 4백 3십만 명에 해당하는 3.5 %로 보고되고 있다. 인구 증가 및 고령화로 인해 이 수치는 2040년까지 1억 2천 2백만 명으로 증가할 것으로 예상한다 (Tham et al., 2014). 특히, 싱가포르 중국인의 경우 약 85%, 미국 흑인 인구의 경우 같은 비율, 미국의 경우 전체 비율이 50%까지도 진단이 어려운 상황이다 (Tatham et al., 2014).

녹내장으로 인한 시력 손실의 대부분은 조기 발견과 이를 통한 적절한 치료를 통해 예방할 수 있으나, 대다수의 녹내장 환자에게 있어서 초기 단계의 녹내장 진단에 실패하고 있다. 그 첫 번째 이유는 녹내장의 질환 특성상 질병의 진행 단계에서 중심 시력이 영향을 받는 만성 녹내장 단계에 이르기까지 환자 대부분은 뚜렷한 자각증상을 느끼지 못하기 때문이다 (Gupta et al., 2016). 녹내장을 초기에 진단하기 어려운 두 번째 이유는 고

도의 전문성과 경험을 보유한 녹내장 전문의만이 안저영상에서 초기 단계 녹내장을 정확하게 판독할 수 있고, 대부분의 녹내장 전문의는 전문 안과 병원 또는 대형 종합병원에만 존재하기 때문이다.

이러한 현상은 녹내장이 초기 단계에서 후기 단계로 진행됨에 따라 치료비가 4배 이상 증가하여 상당한 재정적 부담을 초래하고, 또한 녹내장 말기에 이루어지는 치료 역시 그 예후가 초기 단계부터 치료하는 것보다 상대적으로 좋지 않은 결과를 보인다 (Lee et al., 2006).

국내의 경우 국민건강보험법에 따라서 40세 이상의 모든 국민은 2년마다 의무적으로 건강검진을 받아야 한다. 이 과정에서 검진 대상자는 안저촬영을 선택할 수 있어서 폭넓은 안과 질환의 선별검사 기회가 존재한다. 특히, 녹내장의 경우 이러한 과정에서 선별검사가 이루어져야 하지만, 안저영상에서 망막신경섬유층(Retinal Nerve Fiber Layer, RNFL)의 결손을 육안으로 정확하게 판독하는 것은 고도의 전문성을 요구하는 일이기 때문에 녹내장 조기진단에 실패하는 경우가 많다.

녹내장 발병 초기 단계에서 가장 먼저 결손이 발생하는 해부학적 조직은 신경절세포의 축삭으로 구성된 망막신경섬유층이다. 이러한 병리학적 특성을 고려할 때, 건강검진 또는 안과 검사에서 시행하는 안저촬영을 통해 망막신경섬유층의 결손 여부를 판독해서 조기에 녹내장의 징후를 발견하는 것이 매우 중요하다. 그림1과 그림2는 정상인과 초기 녹내장 환자의 망막을 촬영한 안저영상(Fundus Photograph)이다. 정상인의 안저영상은 그림1에서 화살표로 표시한 부분과 같이 망막혈관 주변에서 시신경 방향으로 주행하는 백발과 같은 많은 망막신경섬유층을 관찰할 수 있다. 초기 녹내장 환자의 경우, 그림2에서 보는 바와 같이 화살표로 표시한 영역에서 국소적인 쐐기모양의 망막신경섬유층 결손을 확인할 수 있다.

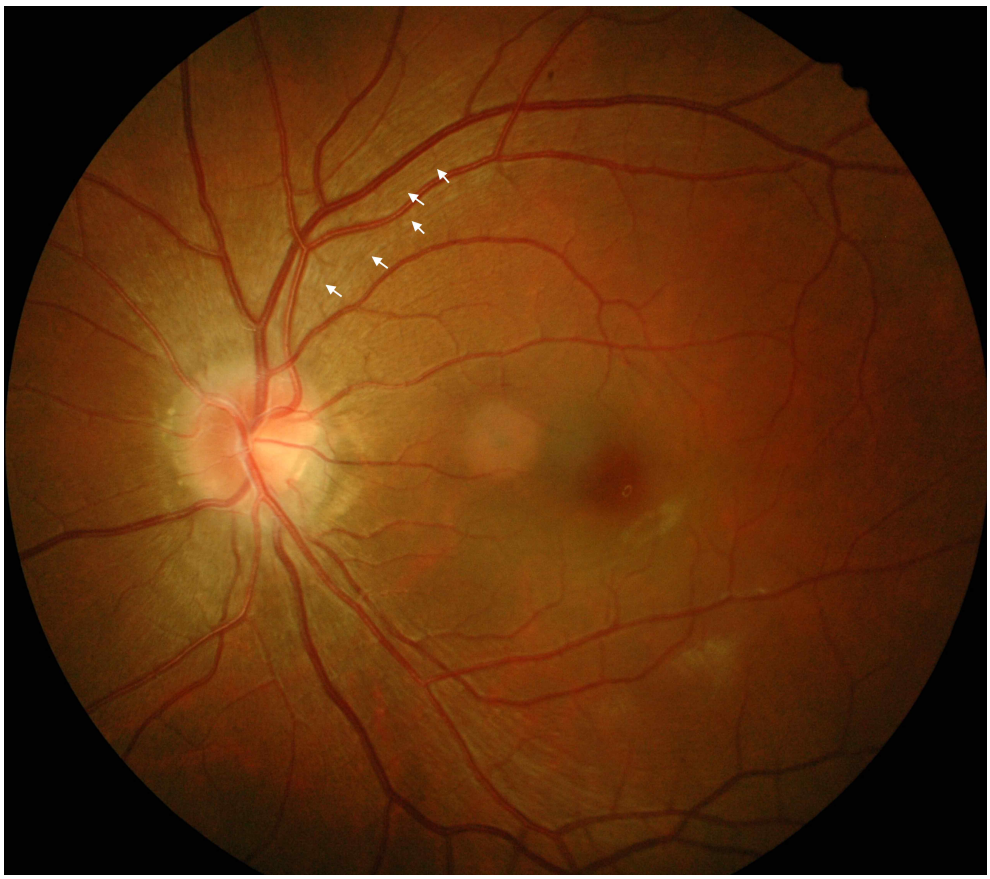


그림 1. 정상(Unaffected Control) 안저 영상

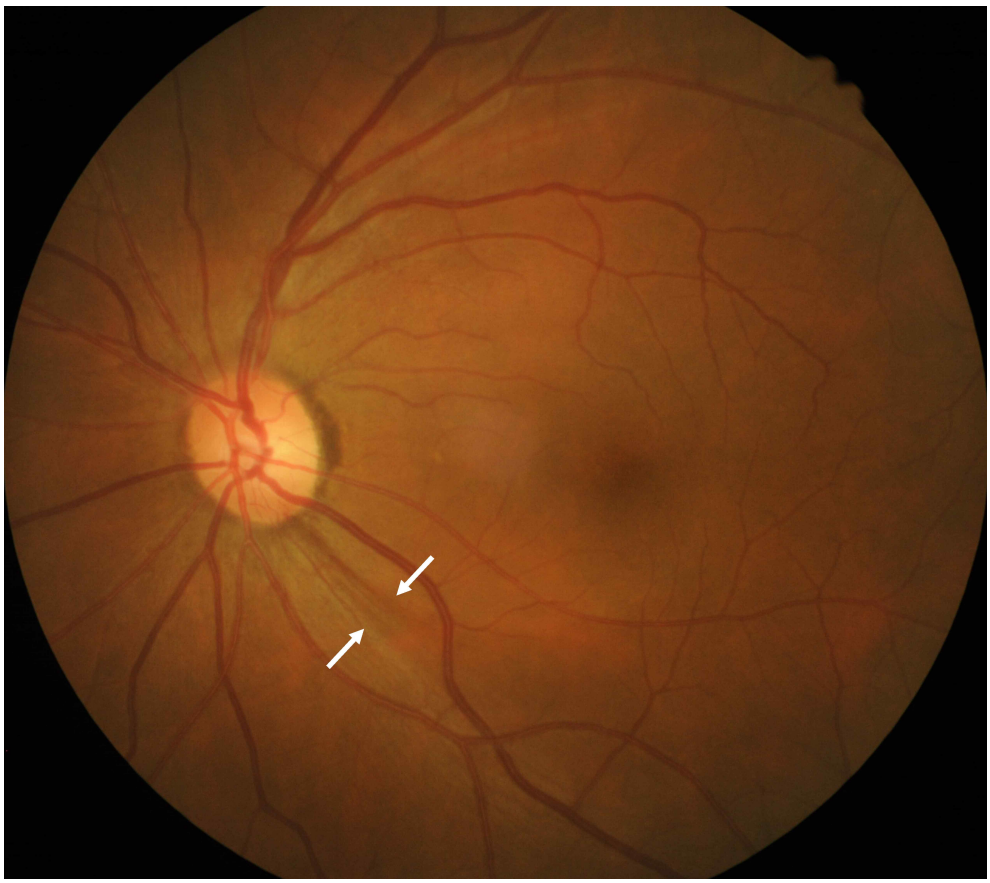


그림 2. 망막신경섬유층 결손이 보이는 안저영상

제 2 절 연구의 필요성

녹내장은 주사레이저현미경(Scanning Laser Polarymetry, SLP) 또는 빛간섭단층촬영(Optical Coherence Tomography, OCT), 시야검사(Visual Field Test), 시신경유두의 함몰 비율 비교가 대표적인 진단방법이다.

OCT를 이용한 진단의 경우, 시신경섬유층의 두께를 측정하는 방식으로 정량화와 객관화 할 수 있는 장점이 있다. 그림3의 OCT 측정 예시에서 보는 바와 같이 OCT 장비는 시신경유두(Optic Disc)를 중심으로 단층촬영 결과를 통해 시신경의 두께를 정량화할 수 있다. 시신경의 두께가 정상인의 평균값보다 낮으면 녹내장을 의심할 수 있는 것이다.

그러나, OCT를 이용한 검사 방법은 시신경유두로부터 매우 제한된 영역에 있는 망막신경섬유층만을 측정하므로 촬영 영역 외부에 존재하는 녹내장과 관련된 병변을 검출하는 데는 한계가 있다. 시야결손전녹내장(Preperimetric Glaucoma) 또는 초기 녹내장의 경우 망막신경섬유층의 결손은 시신경유두에서 시작하여 망막 주변부로 확장되는 양상을 보인다. 따라서, 시신경유두 영역에서 망막신경섬유층 결손이 OCT 측정결과에 유의미하게 나타나지 않을 가능성이 있는 시야결손전녹내장과 초기 녹내장의 진단에는 많은 제약이 있다.

또한, OCT 장비의 경우 전문 안과병원이나 대형병원에서 주로 갖추고 있는 고가의 장비이다. 따라서, 의료 접근성 관점에서 볼 때, 대다수 국민에게 저렴한 비용으로 쉽게 검사할 기회를 제공하지 못한다는 현실적인 제약이 존재한다.

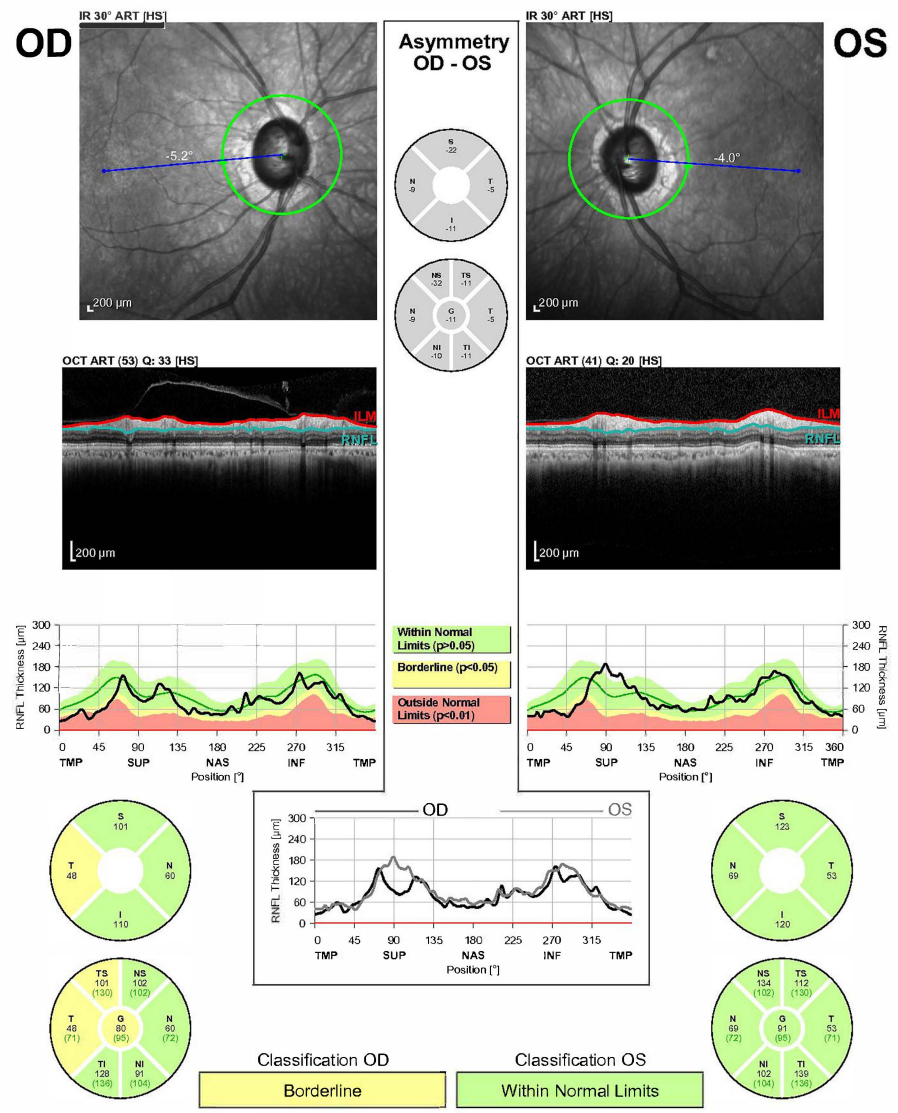


그림 3. OCT 측정 결과 예시

시야 검사는 환자의 미세한 시야 결손 여부를 측정하는 방법으로 녹내장 확진에 많이 사용하고 있다. 그림4에서는 정상인을 대상으로 시야검사를 실시한 결과를 보여주고 있으며, 그림5에서는 녹내장 환자를 대상으로 시야검사 결과를 보여주고 있다. 그림5에서 보는 바와 같이 녹내장 환자의 경우 많은 부분에서 시야 결손이 존재한다는 것을 확인할 수 있다. OCT나 안저 카메라는 환자의 안저에 대한 해부학적 구조 관점에서 측정하는 특징이 있고, 반면에 시야검사는 기능적 관점에 주안점을 두고 측정을 하는 방식이다.

그러나, 시야검사는 측정에 오랜 시간과 노력이 필요하다는 단점과 초기 녹내장의 경우 측정의 일관성이 떨어진다는 단점이 있다. 또 다른 관점에서 시야검사는 시야결손전녹내장의 경우 검진이 어렵다는 단점도 존재한다. 정상에서 초기 녹내장으로 발전하는 중간 단계에 구조적으로는 망막신경섬유층의 결손이 발생했지만, 시야의 기능적 결함이 검출되지 않는 단계가 시야결손전녹내장이다.

또한, OCT와 마찬가지로 시야검사 장비는 전문 안과병원이나 대형병원에서 주로 갖추고 있는 고가의 의료기기이다. 따라서, 의료 접근성 관점에서 볼 때, 대다수 국민에게 저렴한 비용으로 쉽게 검사할 기회를 제공하지 못한다는 현실적인 제약이 존재한다.

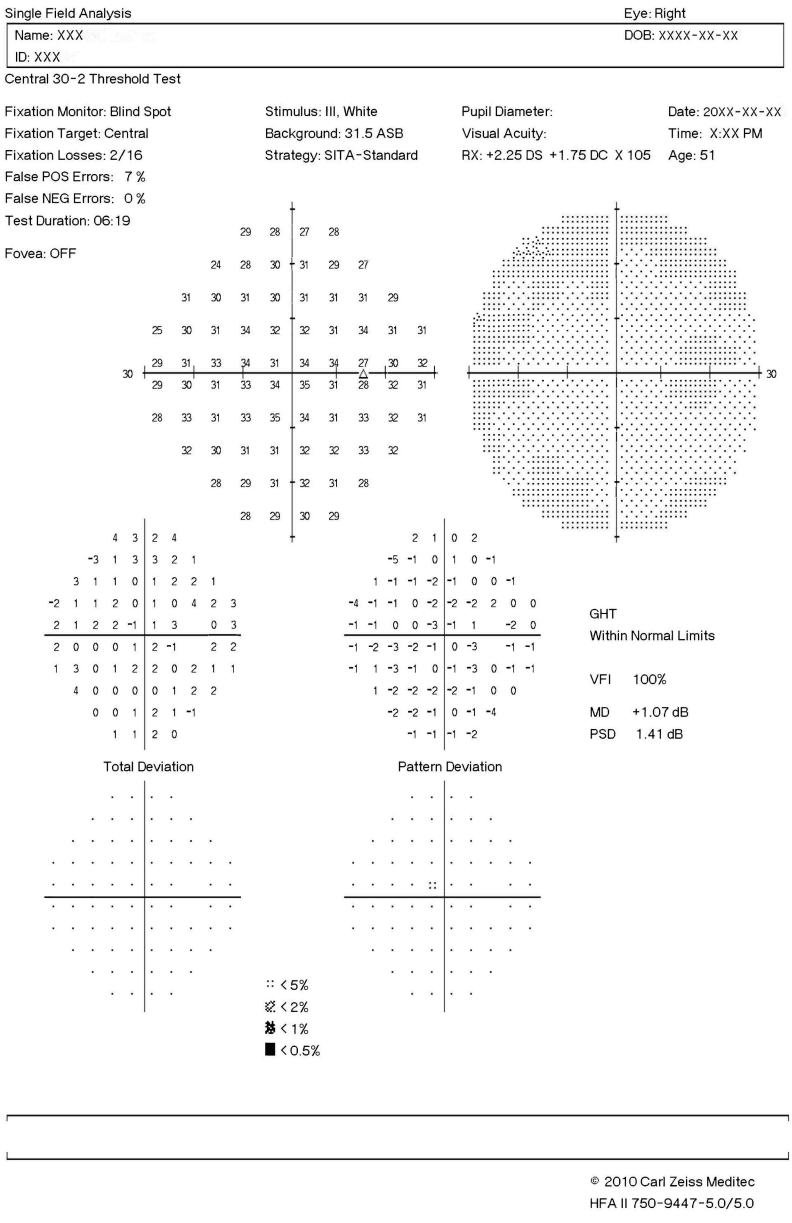


그림 4. 정상인 (Unaffected Control) 시야검사 결과

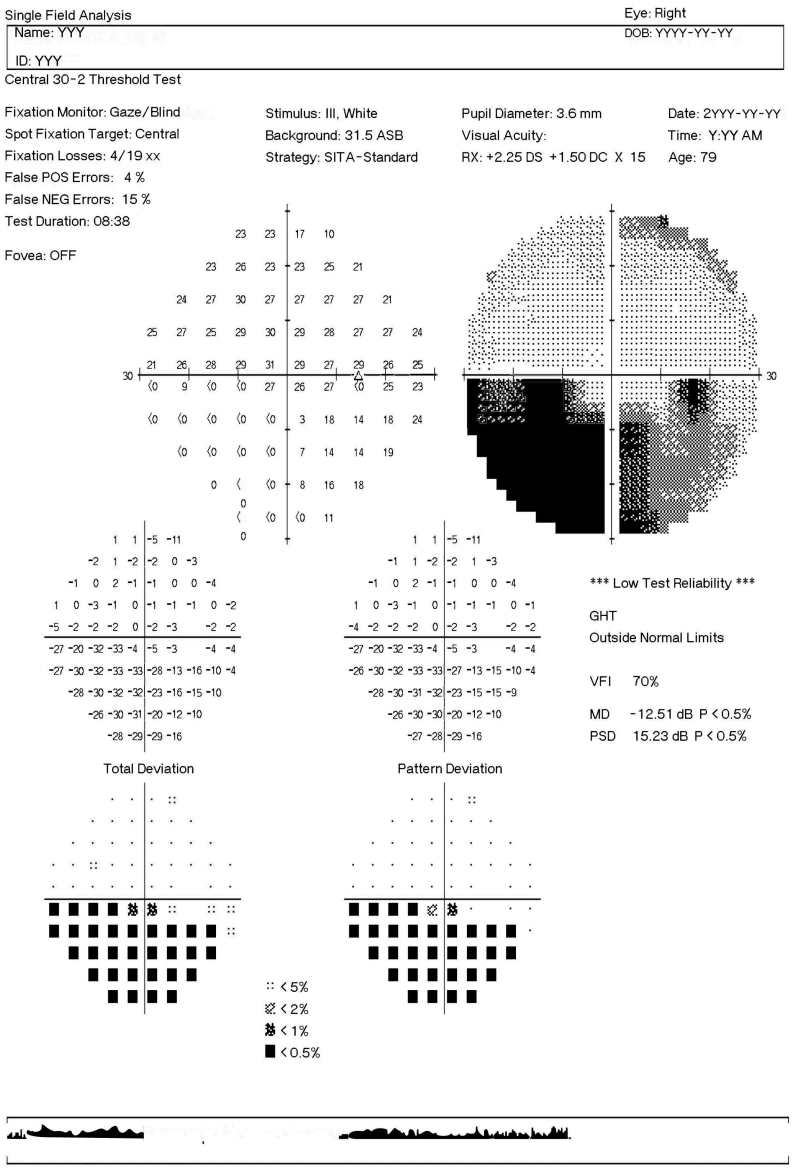


그림 5. 녹내장 환자의 시야검사 결과

안저영상을 이용해 녹내장을 진단하는 종래의 방법은 신경다발을 형성하고 있는 시신경유두에서 유두함몰비(Cup-to-Disc Ratio, CDR)를 계산하는 것이다. CDR 방법 역시 시신경유두라는 매우 제한된 영역을 가지고 진단하기 때문에 시야결손전녹내장과 초기 녹내장을 선별하기에 어려움이 있다. 또한, 망막을 2차원 영상으로 촬영한 결과로는 망막신경섬유층의 두께를 측정할 수 없어서 중증도 정량화 역시 한계가 있다.

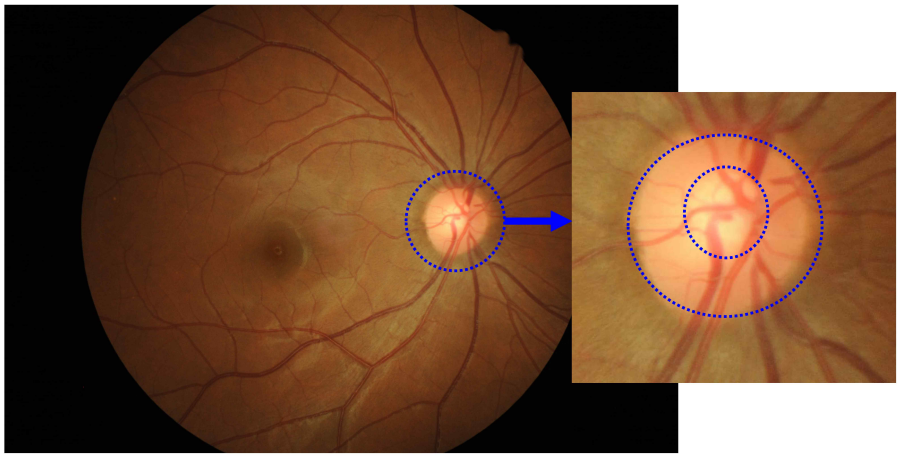


그림 6. CDR 측정 예시

다른 한편으로 녹내장 선별검사는 판독의 정확성과 자동화, 그리고, 미세 변화까지 식별 가능한 검사 해상력이 보장되어야 한다. 이러한 요구조건을 만족하기 위한 기술적 대안으로는 인공지능 분야에서 최근 가장 많은 관심을 받는 딥러닝 기술이 있다. 딥러닝 기술은 Krizhevsky et al. (2012)가 ImageNet에서 적용한 사례를 발표한 이후, 대부분의 인공지능 관련 연구에서는 딥러닝을 기반으로 연구를 진행하고 있다. 다양한 딥러닝 연구결과 중에서 영상 인식에 적합한 방법론이 합성곱신경망(Convolutional Neural Network, CNN)이며, CNN 계열의 대표 연구사례에는 VGGNet (Simonyan et al., 2014), GoogLeNet (Szegedy et al., 2015),

RestNet (He et al., 2016), InceptionNet-v3 (Szegedy et al., 2015), Inception-ResNet-v2 (Szegedy et al., 2017), XceptionNet (Chollet et al., 2017) 등이 있다. 현재까지 합성곱신경망을 기반으로 컴퓨터과학 계열에서 발표된 대부분의 딥러닝 연구는 사물, 동물, 얼굴 등을 인식하는 연구에 집중되어 있다.

녹내장 자동 판독을 위해 딥러닝을 적용한 연구결과로는 Chen et al. (2015)과 Asaoka et al. (2016)가 있다. Chen et al. (2015)은 안저영상과 딥러닝을 접목하여 녹내장을 자동으로 판독하는 연구결과를 발표하였다. Chen et al. (2015)의 연구는 녹내장 판독 정확도 AUROC 0.831 ~ 0.887 수준의 결과를 보였다. 이러한 연구결과는 안저영상에 딥러닝을 적용해서 녹내장을 자동으로 판독할 수 있다는 가능성을 제시한 측면에서 의미가 있다. 그러나, Chen et al. (2015)의 연구는 안저영상에서 시신경유두영역만을 딥러닝에 적용했다. 안저영상의 시신경유두에서 녹내장과 관련된 병변을 식별할 수 있는 상황은 대부분 중기 이상의 녹내장 단계에서 나타나기 때문에 시야결손전녹내장이나 초기 녹내장을 검출하기에는 한계가 있다. 또한, 녹내장 중증도의 등급화에 관련된 논점은 제시하지 못하고 있다.

Li et al. (2018)은 4만 여장의 안저영상으로 InceptionNet-v3을 가지고 학습한 모델로 녹내장을 판독하는 연구결과를 발표하였다. Li et al. (2018)은 4만 여장의 학습 데이터를 가지고 InceptionNet-v3 모델을 단독으로 사용하여 평균 AUROC 0.986, 평균 민감도 95.6%, 평균 특이도 92.0%라는 높은 수준의 성능을 보이는 연구결과를 발표하였다. 그러나, Li et al. (2018)의 연구에서 사용한 안저영상은 안과의사가 레이블링했지만, 시야 검사 결과를 동시에 고려한 레이블링은 하지 않았고, 오로지 안저영상만으로 가지고 레이블링 한 결과를 사용하였다. 이러한 레이블

링 결과는 녹내장을 확진하는 수준의 등급 결정이 아니고, 녹내장 의심 수준의 데이터 레이블 정보를 제공한다는 한계가 있다. 이렇게 레이블링 한 안저영상으로 학습 데이터 집합을 만들고, 딥러닝 모델을 훈련할 경우, 학습한 딥러닝 모델의 결과 역시 녹내장 의심 수준으로 해석해야 하는 한계가 존재하는 것이다.

2016년 동경대에서는 시야 검사 결과를 이용하는 딥러닝 적용사례를 발표하였다 (Asaoka et al., 2016). 시야 검사는 일반적으로 10분 이상의 검사시간이 필요하고, 검사과정에서 환자뿐만 아니라 검사를 주관하는 의료진까지도 많은 노력을 기울여야 하는 단점이 있다. 또한, 시야검사 장비는 안저촬영기보다 상대적으로 고가이기 때문에 의료 현장에 보급된 비율 측면에서 의료 접근성이 좋지 않으며, 저비용으로 녹내장 선별검사를 진행하기에는 어려움이 있다. 또한, 시야 결손이 발생하지 않은 시야결손전녹내장의 검출에는 한계가 있다. Asaoka et al. (2016)의 연구 역시 녹내장 중증도 등급화 쟁점은 고려하지 않았다.

녹내장 선별검사를 자동화하기 위해 현재까지 진행된 연구들을 분석해보면, 안저영상 전체를 이용해 시야결손전녹내장과 초기 녹내장을 자동으로 검출하는 능력을 제공하고, 동시에 녹내장 중증도를 등급화 할 수 있는 기능을 제공하는 딥러닝 연구는 아직 발표된 사례가 없다.

녹내장은 백내장, 황반변성과 더불어 3대 실명 유발 안과 질환으로, 2014년 이루어진 녹내장 질환의 역학조사에 따르면, 2013년 기준으로 전 세계 40대 이상 성인 중에서 약 3.54% (6천 4백 30만명)가 녹내장 환자이며, 2020년에는 그 수가 7천 6백만 명까지 증가할 것으로 추정하고 있다 (Tham et al., 2014). 또한, 2010년 기준으로 전 세계적으로 녹내장으로 실명한 환자는 2백 10만 명이고, 시력을 손실한 환자는 4백 20만 명

에 이른다는 조사 결과가 발표되었다 (Bourne et al., 2017). 경제적 관점에서 녹내장을 분석한 연구결과에 따르면, 녹내장 환자가 부담하는 의료비용은 미국에서만 29억 달러에 이른다고 보고하고 있다 (Varma et al., 2011).

현재 범용으로 사용하고 있는 안저촬영 장비는 녹내장을 초기에 감지할 수 있는 망막신경섬유층의 미세 변화를 자동 판독할 수 있는 기능을 제공하지 못하고 있다. 이와 같은 상황과 녹내장이라는 질병의 특성을 고려해볼 때, 녹내장 진단을 위해 자동화된 선별검사가 가능하고, 녹내장의 중증도를 등급화 할 수 있으며, 시야결손전녹내장과 초기 녹내장 역시 검출 가능한 기술에 대한 연구개발이 절실한 상황이다. 시야결손전녹내장과 초기 녹내장 단계에서부터 발생하는 망막신경섬유층의 결손 여부를 자동으로 판독할 수 있고, 녹내장의 중증도를 등급화 할 수 있는 기술을 개발한다면, 향후 의료서비스 품질을 향상할 수 있고, 관련 의료비용 지출의 절감과 함께 실명 위험률을 낮춤으로써 실명에 따른 사회적 비용 역시 절감하는 효과를 기대할 수 있다.

따라서, 본 연구에서는 진단 검사 비용이 저렴하면서 동시에 의료접근성이 좋은 안저영상을 바탕으로 시야결손전녹내장과 초기 녹내장의 선별검사가 가능하며, 과거 검사 이력과 비교를 통해 녹내장 진행정도를 쉽게 확인 할 수 있는 딥러닝 모델을 연구하고자 한다. 향후, 본 연구를 통해 개발한 연구결과를 의료장비에 탑재하여 건강검진과 안과 진단에 활용한다면, 의료기기 시장에 미치는 파급효과는 매우 클 것으로 예상하며, 이에 대한 선제적 대응 차원에서도 본 연구의 진행은 필수적이라 할 수 있다.

제 3 절 연구의 목표 및 내용

본 연구의 최종목표는 안저영상에 나타난 망막신경섬유층 결손 여부를 자동으로 판독할 수 있고, 이를 바탕으로 녹내장의 중증도를 등급화 할 수 있는 딥러닝 모델을 개발하는 것이다. 이러한 딥러닝 모델을 연구하는 목적은 시야결손전녹내장과 초기 녹내장을 선별검사 할 수 있고, 녹내장 중증도를 객관적으로 등급화 할 수 있는 모델을 의료 현장에 제공하기 위함이다. 의료 현장에 이러한 인공지능 기술을 기반으로 하는 도구를 제공함으로써 환자뿐만 아니라 의료진에게도 경제적, 사회적 이득을 제공하는 것이다.

녹내장 조기진단을 위한 딥러닝 모델을 개발하기 위해서는 3가지 세부 목표를 유기적으로 연계해서 연구를 진행해야 한다. 3가지 세부 목표에는 딥러닝 기계학습을 위한 안저영상 DB 구축, 안저영상 전처리 및 후처리 알고리즘 개발, 녹내장 진단 딥러닝 모델 개발이 있다.

본 연구에서는 안저영상을 바탕으로 녹내장을 선별 검사할 수 있고, 녹내장의 중증도를 등급화 할 수 있는 딥러닝 모델을 개발하였다. 특히, 녹내장 진단을 위한 딥러닝 모델 개발을 위해 영상 인식 분야에 특화된 합성곱신경망을 바탕으로 하였으며, 다양한 형태와 특성을 갖는 여러 개의 합성곱신경망 모델을 데이터 교차검증이 완료된 고품질의 안저영상을 이용해 훈련하였다. 훈련이 완료된 여러 개의 합성곱신경망 모델을 본 연구에서 제안하는 앙상블 방법을 통해 융합하였고, 기존에 발표된 방법과 성능을 비교하였다.

제 2 장 연구 방법

안저영상을 바탕으로 녹내장 진단을 위한 딥러닝 모델을 개발하기 위해서는 가장 먼저 환자로부터 촬영한 안저영상을 모아 정리하고, 데이터베이스를 구축하는 것이다. 안저영상 데이터베이스는 안저영상의 전처리 및 후처리를 완료한 데이터를 저장하는 곳이다.

안저영상의 데이터베이스 구축을 완료한 다음, 딥러닝 구조에 데이터를 입력하고, 실험을 진행하기 위한 실험계획을 수립한다. 실험계획의 수립에 있어서 가장 중요한 요소는 데이터와 연구의 특성을 자세히 파악하여 당초 계획한 연구 목표를 달성 할 수 있는 최적의 실험계획을 도출하는 것이다.

딥러닝을 위한 데이터 준비와 실험계획 수립이 완료되면, 설계한 딥러닝 학습 실험계획에 따라 딥러닝 기계학습 실험을 진행한다. 모든 실험의 결과는 매번 일정한 기준으로 그 결과를 평가하고, 고찰을 통해 성능향상 기법을 적용할 수 있는 실험계획을 수정하여 다시 기계학습 실험을 반복적으로 수행한다. 이러한 반복 실험과정을 통해 최적의 성능을 보이는 녹내장 진단 모델을 탐색하는 것이 본 연구 진행의 전반적인 절차이다.

제 1 절 안저영상 데이터베이스 구축

1. 데이터 측정 환경 및 방법

본 연구에서 사용한 안저영상은 동공 확장 없이 디지털 안저 카메라(Nonmyd 7, Kowa Optimed, Tokyo, Japan)를 사용하여 컬러 영상으로 촬

영한 결과를 사용하였다. 촬영 과정에서 시신경을 중심으로 망막에 초점을 맞추었으며, 촬영에서 획득한 컬러 영상을 RNFL 사진(Red-Free 영상, Green Channel 데이터)으로 변환하여 컬러 안저영상과 RNFL 안저영상 모두를 LCD 모니터에서 검토하였다.

촬영한 안저영상의 레이블링 정확도를 높이기 위해 별도의 시야검사 데이터를 측정하여 최종 레이블링 작업에 같이 사용하였다. 시야검사는 Zeiss 사의 험프리 자동시야계, SITA strategy를 시행했다.

녹내장은 안저 사진에서 명백한 시신경유두 및 망막신경섬유층의 결손과 그에 상응하는 시야 결손이 확인된 경우로 정의하였다. 녹내장성 시신경유두 함몰비의 증가, 시신경테 (Neuroretinal rim) 패임, 망막신경섬유층의 결손으로 정의되었다. 녹내장성 시야 결손은 패턴 편차 도표에서 3개 이상의 인접한 점들의 역치가 정상 집단의 5% 미만으로 ($P < 5\%$) 나타나고 그 중 한점이 1% 미만일 때 이거나, 패턴 표준 편차 (Pattern standard deviation) 지표가 5% 미만으로 나타날 때로 정의하였다. 주시상실이 20% 미만이고 위양성이 33% 미만일 때 시야 검사를 신뢰할만한 것으로 정의하였고, 본 연구에서는 신뢰할 수 있는 데이터만 선별하여 딥러닝 모델에 사용하였다.

2. 안저영상 레이블링 방법

딥러닝 기계학습의 입력에 사용하는 모든 안저영상에 대해서 해당 영상에 맞는 녹내장 등급 정보를 할당해야 한다. 이러한 작업을 레이블링이라 한다. 녹내장 등급의 결정은 대한민국 녹내장 학회 정회원 자격을 갖춘 4명의 안과 전문의가 안저영상과 함께 OCT 측정결과와 시야검사 결과를 동시에 검토해서 녹내장 등급을 결정하였다.

녹내장의 등급은 다음 표와 같은 기준으로 정의하였다. 본 연구에서 정의한 녹내장 등급의 주요한 특징으로는 정상과 초기 녹내장 사이에 시야결손전녹내장(preperimetric)을 별도의 등급으로 정의한 것이다. 시야결손전녹내장(preperimetric)의 판단은 시야검사 결과와 OCT 촬영 결과를 동시에 고려해서 결정하였다.

정상 등급은 시야검사와 안저영상에서 특이한 질환의 특징이 없는 것으로 정의하였고, 녹내장 전기(G0) 등급은 시야검사에서는 정상이나 안저영상과 OCT 판독에서 망막신경섬유층 결손이 보이는 경우로 정의하였다. 녹내장 초기 등급(G1)은 시야검사 기준값(MD, Mean Deviation)이 0보다 작고 -6.0보다 큰 경우로 하였으며, 녹내장 중기 등급(G2)은 시야검사 기준값이 -6.0보다 작고, -12.0보다 큰 경우로 하였고, 마지막으로 녹내장 말기 등급(G3)은 시야검사 기준값이 -12.0보다 작은 경우로 정의하였다.

디옵터 -3.0D보다 작거나 3.0D 이상인 난시, 녹내장 평가를 저해할 수 있는 안저영상의 품질이 낮은 영상은 학습용 데이터에서 제외하였다. 또한, 염증, 허혈, 압축, 경색 등으로 유발된 다른 시신경 병증이 있는 피험자의 영상과 망막 박리, 나이 관련 황반변성, 당뇨 망막 병증, 망막 혈관 폐쇄 등과 같은 질환을 동반한 피험자의 안저영상 역시 학습 데이터에서 제외하였다.

일차적으로 등급이 결정된 안저영상 중에서 정상, 시야결손전녹내장, 초기 녹내장 영상에 대해서는 안과 전문의의 교차검증을 통해 만장일치 의견으로 등급이 결정된 것만을 본 연구의 기계학습 데이터에 포함하였고, 의견일치가 이루어지지 않은 안저영상은 기계학습 데이터에서 제외하였다.

표 1. 녹내장 등급 및 분류별 클래스 코드

등급	녹내장 등급 코드	녹내장 선별검사 클래스 코드 (이진 분류)	녹내장 중증도 등급화 클래스 코드 (삼진 분류)	비고
Unaffected Control (정상)	NN	C0	C0	
Preperimetric Glaucoma Grade (녹내장 전기)	G0	C1	C1	시야검사는 정상이나 안저영상과 OCT 판독에서 RNFL 결손이 보이는 경우
Mild Glaucoma Grade (녹내장 초기)	G1			-6.0 < MD < 0.0
Moderate Glaucoma Grade (녹내장 중기)	G2		C2	-12.0 < MD <= -6.0
Severe Glaucoma Grade (녹내장 말기)	G3			MD <= -12.0

※ MD: 험프리 시야검사 결과의 Mean Deviation

3. 안저영상 고유 식별자 할당 방법

녹내장 등급 결정된 안저영상에 대해서는 환자 번호, 촬영일, 좌측안, 우측안 정보를 코드화하여 영상의 고유 식별자로 정의하였다. 모든 영상마다 이처럼 고유 식별자 정보를 부여한 이유는 학습결과에 왜곡이 발생하는 것을 방지하기 위해서 환자 1명당 최대 2장(좌측안 1장, 우측안 1장) 이내로 기계학습 데이터를 구성하기 위함이며, 기계학습과정에서 대량의 데이터를 일관성 있게 관리하기 위함이다.

제 2 절 안저영상 처리 방법

1. 안저영상 처리 개요

안저영상마다 녹내장 등급과 고유 식별자를 할당한 다음에는 각각의 영상마다 전처리 작업과 후처리 작업을 수행하였다. 전처리 작업에서는 영상에 포함된 잡음과 개인 식별정보를 제거하는 것이고, 후처리 작업에서는 딥러닝 기계학습 효과를 극대화하기 위한 영상 증폭 및 변환을 수행하는 것이며, 전반적인 절차는 아래 그림과 같다.

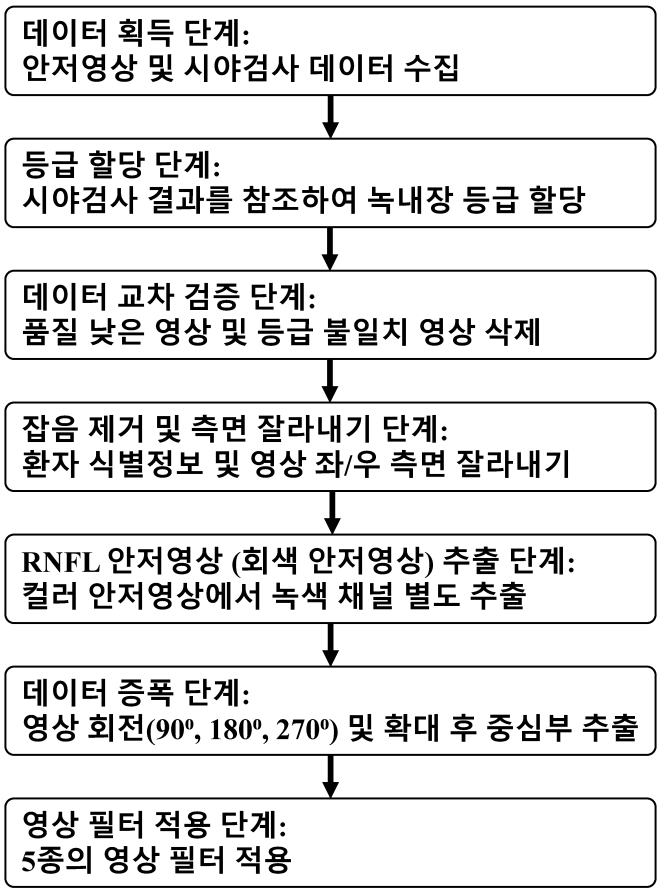


그림 7. 딥러닝 모델 학습용 데이터 구축 절차

2. 안저영상 전처리 방법

촬영한 안저영상 원본은 영상의 좌, 우측에 흑색 영역이 존재하며, 이는 영상 판독에 어떤 영향도 주지 않는다. 따라서 영상에서 흑색 부분을 제거하는 작업은 기계학습에서 컴퓨팅 파워를 절약할 수 있으며, 필요 없는 데이터를 제거함으로써 학습효과를 높일 수 있다. 또한, 일부 원본 영상에는 환자의 개인 식별 정보 또는 진료과정에서 인위적으로 기록한 정보들이 표시되어 있기도 하다. 이러한 정보 역시 정상적인 기계학습을 위해 반드시 제거해야 하는 영역이다. 이러한 잡음영역과 개인 식별 정보를 제거하는 영상 전처리 작업의 개념을 그림8과 같이 도식화 하였다.

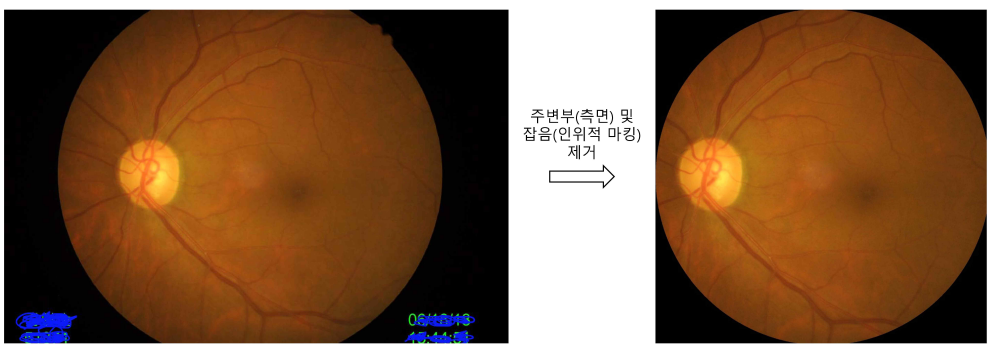


그림 8. 잡음 및 개인 식별 정보 제거 개념도

3. 안저영상 후처리 방법

기계학습에 활용성을 극대화하기 위해 전처리가 완료된 영상을 대상으로 회전, 확대 후 추출을 수행하는 데이터 증폭 작업을 진행하였다. 또한, 질환 판독의 성능을 극대화하기 위해 다양한 종류의 영상 필터를 적용하여 영상 변환을 시행하였다.

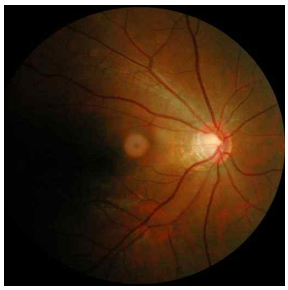


그림 9. 원본 영상

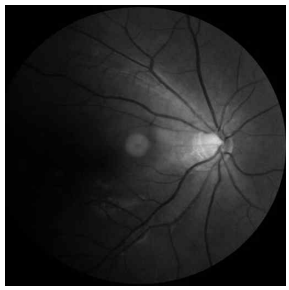


그림 10. 회색 변환 영상

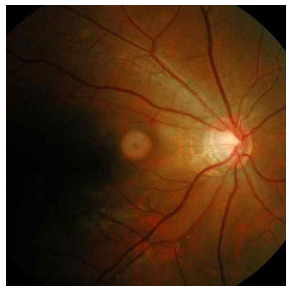


그림 11. 확대 후 추출 영상

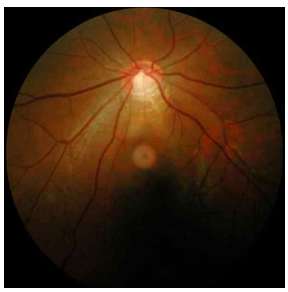


그림 12. 90도 회전 영상

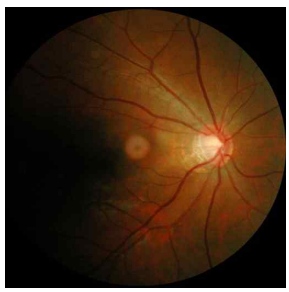


그림 13. Bilateral 필터 적용 영상

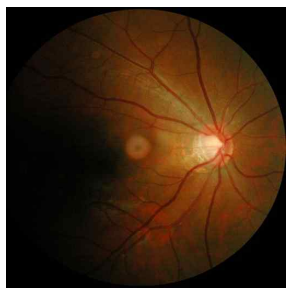


그림 14. Gaussian 필터 적용 영상

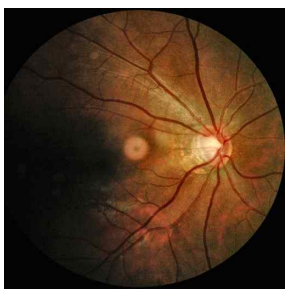


그림 15. Histogram Equalization 필터 적용 영상

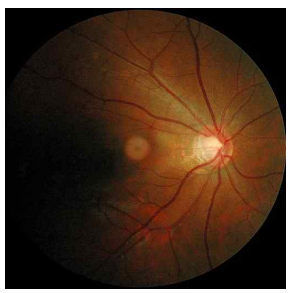


그림 16. Sharpening 필터 적용 영상

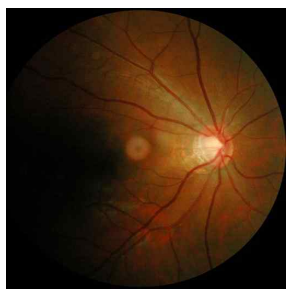


그림 17. Median 필터 적용 영상

제 3 절 합성곱신경망 모델

1. InceptionNet-v3과 Inception-ResNet-v2 모델

2012년 Krizhevsky et al. (2012)의 딥러닝이 발표된 이후로 다양한 딥러닝 연구결과들이 발표되었다. 다양한 딥러닝 결과 중에 최근에 주목받고 있는 대표적인 딥러닝 모델로는 InceptionNet-v3와 Inception-ResNet-v2가 있다.

InceptionNet-v3은 Szegedy et al. (2015)가 발표한 합성곱신경망 방법으로, 여러 개의 인셉션(inception) 모듈을 핵심 구성요소로 하는 합성곱신경망이다. 인셉션 모듈은 아래 그림에서 보는 바와 같이 합성곱신경망의 계층과 계층 사이를 연결하는 역할을 하며, 인셉션 모듈 내부는 다양한 컨볼루션 연산을 수행하는 마이크로 구조들을 결합하여 작동한다. InceptionNet-v3는 Li et al. (2018)이 녹내장 진단용 딥러닝 모델에서도 채용해서 사용했다.

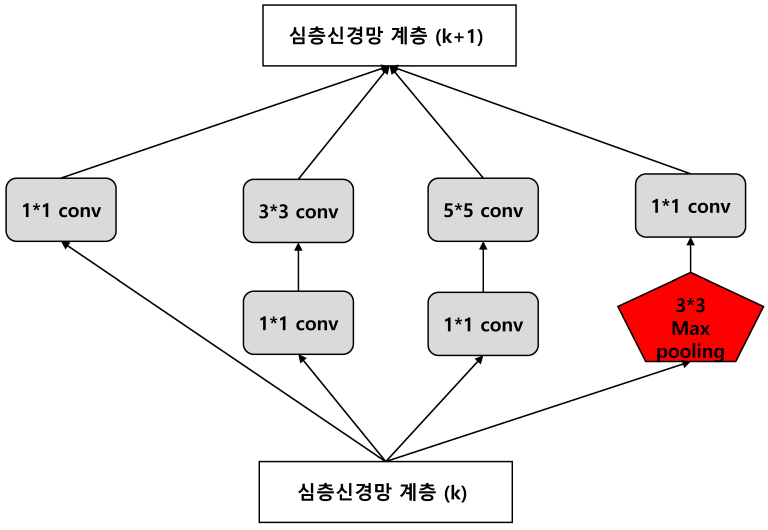


그림 18. Inception 모듈 구조도

다양한 딥러닝 모델이 있지만, 대부분의 딥러닝 모델들은 사물, 동물 등을 인식하기 위해 연구된 결과들이기 때문에 의료분야에 적용하기 위해서는 기존의 모델을 일부 수정해서 적용해야 한다. 본 연구에서는 Li et al. (2018)에서 사용한 InceptionNet-v3를 기준 모델로 하여 본 연구에서 제시한 앙상블 방법과 비교하였다. 모델 앙상블을 적용하기 위해서 InceptionNet-v3와 Inception-ResNet-v2의 출력단을 녹내장 판독에 맞게 다양한 변화를 적용하여 판독 모델을 학습시켰다.

2. 정상과 녹내장 선별을 위한 이진 분류 CNN 모델

이진 분류는 예측 모델의 최종 출력이 2개이고, 통상적으로 정상 또는 비정상을 분류하는 목적에 많이 사용한다. 본 연구에서 정상과 녹내장을 구분하기 위한 목적으로 이진 분류 형식을 갖추어 아래 그림 19과 같이 InceptionNet-v3와 Inception-ResNet-v2를 녹내장 진단에 맞게 변형하여 사용하였다.

3. 녹내장 중등도 등급화를 위한 삼진 분류 CNN 모델

삼진 분류는 예측 모델의 최종 출력이 3개이고, 3개의 클래스를 갖는 모델에 적용한다. 본 연구에서는 정상, 초기 녹내장, 중증 녹내장으로 등급화하기 위한 목적으로 삼진 분류 형식을 갖추어 아래 그림 20과 같이 InceptionNet-v3과 Inception-ResNet-v2를 녹내장 중증도 등급화에 맞게 변형하여 사용하였다.

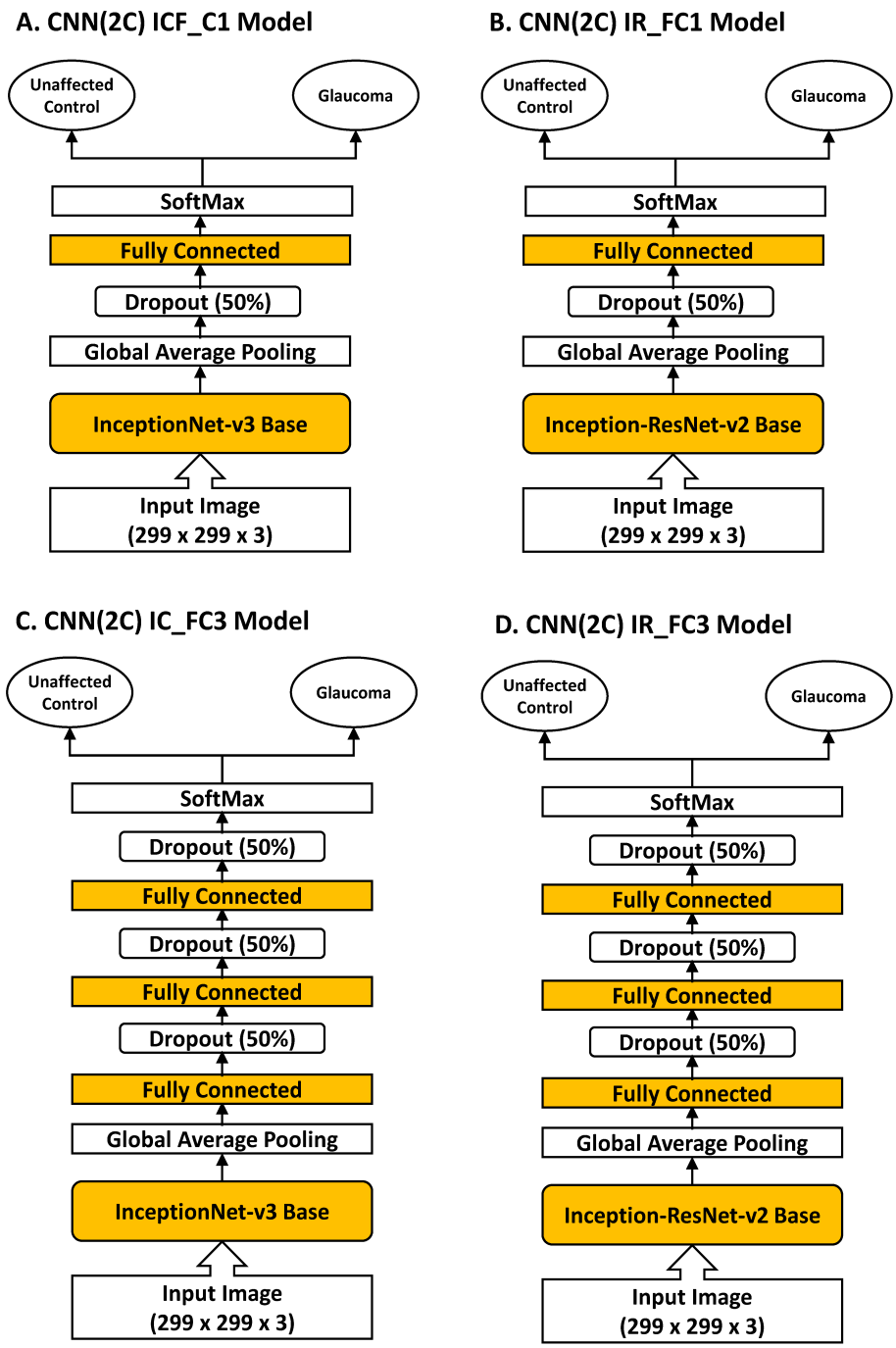


그림 19. 녹내장 선별검사 CNN 구조도

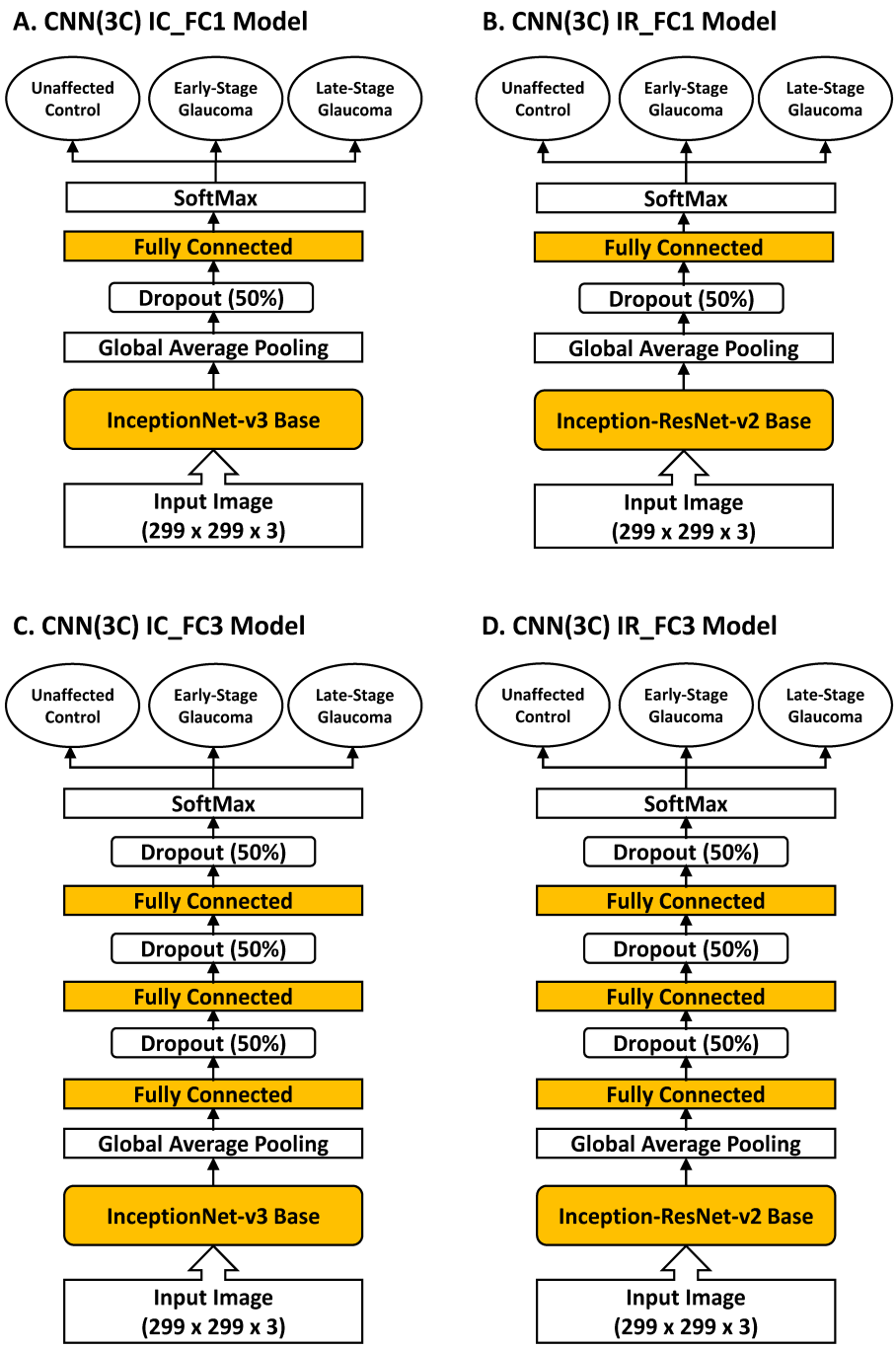


그림 20. 녹내장 중증도 등급화 CNN 구조도

제 4 절 기계학습 실험 설계 및 앙상블 전략

안저영상을 사용해서 녹내장 진단 및 중증도 등급화를 목적으로 하는 합성곱신경망을 학습시키기 위해서는 다양한 요인변수에 대한 최적값을 탐색하는 실험을 진행해야 한다. 합성곱신경망의 성능에 영향을 주는 요인변수로는 입력 영상의 크기, 학습 데이터 규모, 입력 데이터 특성 등이 있다.

표 2. 합성곱신경망 실험 요인변수 코드표

기준	코드	설명
영상 필터 종류	B	Bilateral 필터 적용 영상 사용
	G	Gaussian 필터 적용 영상 사용
	H	Histogram Equalization 필터 적용 영상 사용
	M	Median 필터 적용 영상 사용
	O	원본 영상 그대로 사용
	S	Sharpening 필터 적용 영상 사용
영상 확대 여부	O	원본 영상 그대로 사용
	U	원본을 125% 확대한 후 중심점을 중심으로 299x299 크기로 잘라낸 영상 사용
영상 회전 여부	O	원본 영상 그대로 사용
	R	원본을 90, 180, 270도 회전한 영상 사용
영상 컬러	C	컬러 영상 사용
	G	회색 영상 사용
CNN 모델	IC	InceptionNet-v3 모델 정의 사용
	IR	Inception-ResNet-v2 모델 정의 사용
Fully connected layer	FC1	1개 층의 fully connected layer 사용
	FC3	3개 층의 fully connected layers 사용

이러한 변수 중에 입력 영상의 크기는 CNN 모델 정의에 따라 자동으로 결정되는 변수이며, 본 연구에서는 InceptionNet-v3과 Inception-ResNet-v2를 사용하기 때문에 299x299 크기의 3채널의 영상으로 하였다. 나머지 요인 변수는 다양한 조건의 조합으로 수많은 개별 실험들을 수행할 수 있으므로, 전체적인 실험 결과의 일관성 있는 진행 및 결과 평가 및 정리를 위해 표2와 같은 구조로 개별 실험을 정의하였다.

표2와 같이 정의한 실험 요인변수를 최대로 조합할 경우, 모든 필터를 적용하고, 모든 데이터 증복 방법을 적용하면 BGHMOS-OU-OR-CG-XX와 같은 형태의 실험을 구성할 수 있다. 본 연구에서는 녹내장 선별검사 및 녹내장 중증도 등급화를 위해 표3과 표4에 정의한 실험을 생성하고 앙상블 방법을 적용하였다.

표3과 표4에서 제시한 48개의 개별 CNN 모델은 5개의 구성요소를 결합하여 식별자를 부여하였다. 5개의 구성요소로는 모델에서 판독 클래스의 종류, 영상 필터, 영상 컬러 형태, CNN 기반 모델, fully connected layer의 계층 수량이다. “CNN(2C)_B_C_IC_FC1” 을 사례로 모델 식별자 구성 규칙을 설명하면, 모델의 최종 출력이 2개인 CNN 모델이며, 이것은 녹내장 선별검사를 목적으로 하는 이진 분류를 의미하고, Bilateral filter를 사용한 컬러 영상을 훈련 및 검증에 사용하였으며, InceptionNet-v3를 기반으로 fully connected layer 1계층을 사용한 모델이라는 의미이다. “CNN(3C)_” 로 시작하는 모델은 최종 출력이 3개인 CNN 모델이며, 녹내장 중증도 등급화를 목적으로 하는 삼진 분류를 의미하고, 나머지 해석은 이진 분류와 같다. 또한, 영상의 회전과 확대 후 추출을 모든 개별 모델의 데이터 증폭 방법으로 동일하게 적용하였다.

표 3. 녹내장 선별검사 앙상블에 포함한 개별 모델

No	Model Code	No	Model Code
1	CNN(2C)_B_C_IC_FC1	25	CNN(2C)_M_C_IC_FC1
2	CNN(2C)_B_C_IC_FC3	26	CNN(2C)_M_C_IC_FC3
3	CNN(2C)_B_C_IR_FC1	27	CNN(2C)_M_C_IR_FC1
4	CNN(2C)_B_C_IR_FC3	28	CNN(2C)_M_C_IR_FC3
5	CNN(2C)_B_G_IC_FC1	29	CNN(2C)_M_G_IC_FC1
6	CNN(2C)_B_G_IC_FC3	30	CNN(2C)_M_G_IC_FC3
7	CNN(2C)_B_G_IR_FC1	31	CNN(2C)_M_G_IR_FC1
8	CNN(2C)_B_G_IR_FC3	32	CNN(2C)_M_G_IR_FC3
9	CNN(2C)_G_C_IC_FC1	33	CNN(2C)_O_C_IC_FC1
10	CNN(2C)_G_C_IC_FC3	34	CNN(2C)_O_C_IC_FC3
11	CNN(2C)_G_C_IR_FC1	35	CNN(2C)_O_C_IR_FC1
12	CNN(2C)_G_C_IR_FC3	36	CNN(2C)_O_C_IR_FC3
13	CNN(2C)_G_G_IC_FC1	37	CNN(2C)_O_G_IC_FC1
14	CNN(2C)_G_G_IC_FC3	38	CNN(2C)_O_G_IC_FC3
15	CNN(2C)_G_G_IR_FC1	39	CNN(2C)_O_G_IR_FC1
16	CNN(2C)_G_G_IR_FC3	40	CNN(2C)_O_G_IR_FC3
17	CNN(2C)_H_C_IC_FC1	41	CNN(2C)_S_C_IC_FC1
18	CNN(2C)_H_C_IC_FC3	42	CNN(2C)_S_C_IC_FC3
19	CNN(2C)_H_C_IR_FC1	43	CNN(2C)_S_C_IR_FC1
20	CNN(2C)_H_C_IR_FC3	44	CNN(2C)_S_C_IR_FC3
21	CNN(2C)_H_G_IC_FC1	45	CNN(2C)_S_G_IC_FC1
22	CNN(2C)_H_G_IC_FC3	46	CNN(2C)_S_G_IC_FC3
23	CNN(2C)_H_G_IR_FC1	47	CNN(2C)_S_G_IR_FC1
24	CNN(2C)_H_G_IR_FC3	48	CNN(2C)_S_G_IR_FC3

- ❍ **Model Coding Rule**: ImageFilter_ImageColor_CNNBase_FullyConnectedLayer
- ❍ **Image Filter Code**: B=Bilateral Filter, G=Gaussian Filter, H=Histogram Equalization Filter, M=Median Filter, O=No Filter Used, S=Sharpening Filter
- ❍ **Image Color Code**: C=Color fundus photographs. G=RNFL photographs
- ❍ **CNN Base Code**: IC=InceptionNet-v3, IR=Inception-ResNet-v2
- ❍ **Fully Connected Layer Code**: FC1=1 Fully Connected Layer, FC3=3 Fully Connected Layers

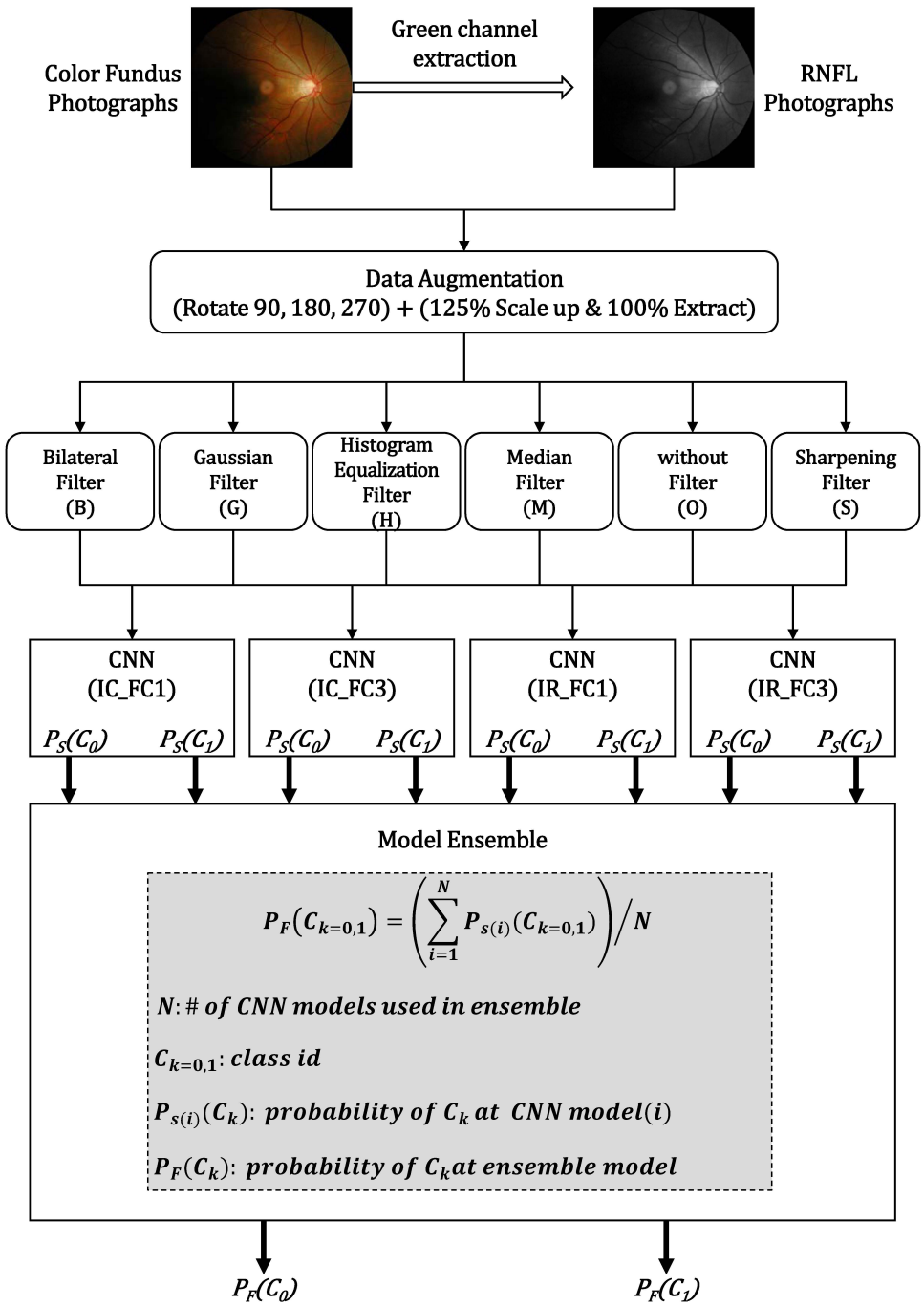


그림 21. 녹내장 선별검사 앙상블 개념도

표 4. 녹내장 중증도 등급화 앙상블에 포함한 개별 모델

No	Model Code	No	Model Code
1	CNN(3C)_B_C_IC_FC1	25	CNN(3C)_M_C_IC_FC1
2	CNN(3C)_B_C_IC_FC3	26	CNN(3C)_M_C_IC_FC3
3	CNN(3C)_B_C_IR_FC1	27	CNN(3C)_M_C_IR_FC1
4	CNN(3C)_B_C_IR_FC3	28	CNN(3C)_M_C_IR_FC3
5	CNN(3C)_B_G_IC_FC1	29	CNN(3C)_M_G_IC_FC1
6	CNN(3C)_B_G_IC_FC3	30	CNN(3C)_M_G_IC_FC3
7	CNN(3C)_B_G_IR_FC1	31	CNN(3C)_M_G_IR_FC1
8	CNN(3C)_B_G_IR_FC3	32	CNN(3C)_M_G_IR_FC3
9	CNN(3C)_G_C_IC_FC1	33	CNN(3C)_O_C_IC_FC1
10	CNN(3C)_G_C_IC_FC3	34	CNN(3C)_O_C_IC_FC3
11	CNN(3C)_G_C_IR_FC1	35	CNN(3C)_O_C_IR_FC1
12	CNN(3C)_G_C_IR_FC3	36	CNN(3C)_O_C_IR_FC3
13	CNN(3C)_G_G_IC_FC1	37	CNN(3C)_O_G_IC_FC1
14	CNN(3C)_G_G_IC_FC3	38	CNN(3C)_O_G_IC_FC3
15	CNN(3C)_G_G_IR_FC1	39	CNN(3C)_O_G_IR_FC1
16	CNN(3C)_G_G_IR_FC3	40	CNN(3C)_O_G_IR_FC3
17	CNN(3C)_H_C_IC_FC1	41	CNN(3C)_S_C_IC_FC1
18	CNN(3C)_H_C_IC_FC3	42	CNN(3C)_S_C_IC_FC3
19	CNN(3C)_H_C_IR_FC1	43	CNN(3C)_S_C_IR_FC1
20	CNN(3C)_H_C_IR_FC3	44	CNN(3C)_S_C_IR_FC3
21	CNN(3C)_H_G_IC_FC1	45	CNN(3C)_S_G_IC_FC1
22	CNN(3C)_H_G_IC_FC3	46	CNN(3C)_S_G_IC_FC3
23	CNN(3C)_H_G_IR_FC1	47	CNN(3C)_S_G_IR_FC1
24	CNN(3C)_H_G_IR_FC3	48	CNN(3C)_S_G_IR_FC3

- ❍ **Model Coding Rule**: ImageFilter_ImageColor_CNNBase_FullyConnectedLayer
- ❍ **Image Filter Code**: B=Bilateral Filter, G=Gaussian Filter, H=Histogram Equalization Filter, M=Median Filter, O=No Filter Used, S=Sharpening Filter
- ❍ **Image Color Code**: C=Color fundus photographs. G=RNFL photographs
- ❍ **CNN Base Code**: IC=InceptionNet-v3, IR=Inception-ResNet-v2
- ❍ **Fully Connected Layer Code**: FC1=1 Fully Connected Layer, FC3=3 Fully Connected Layers

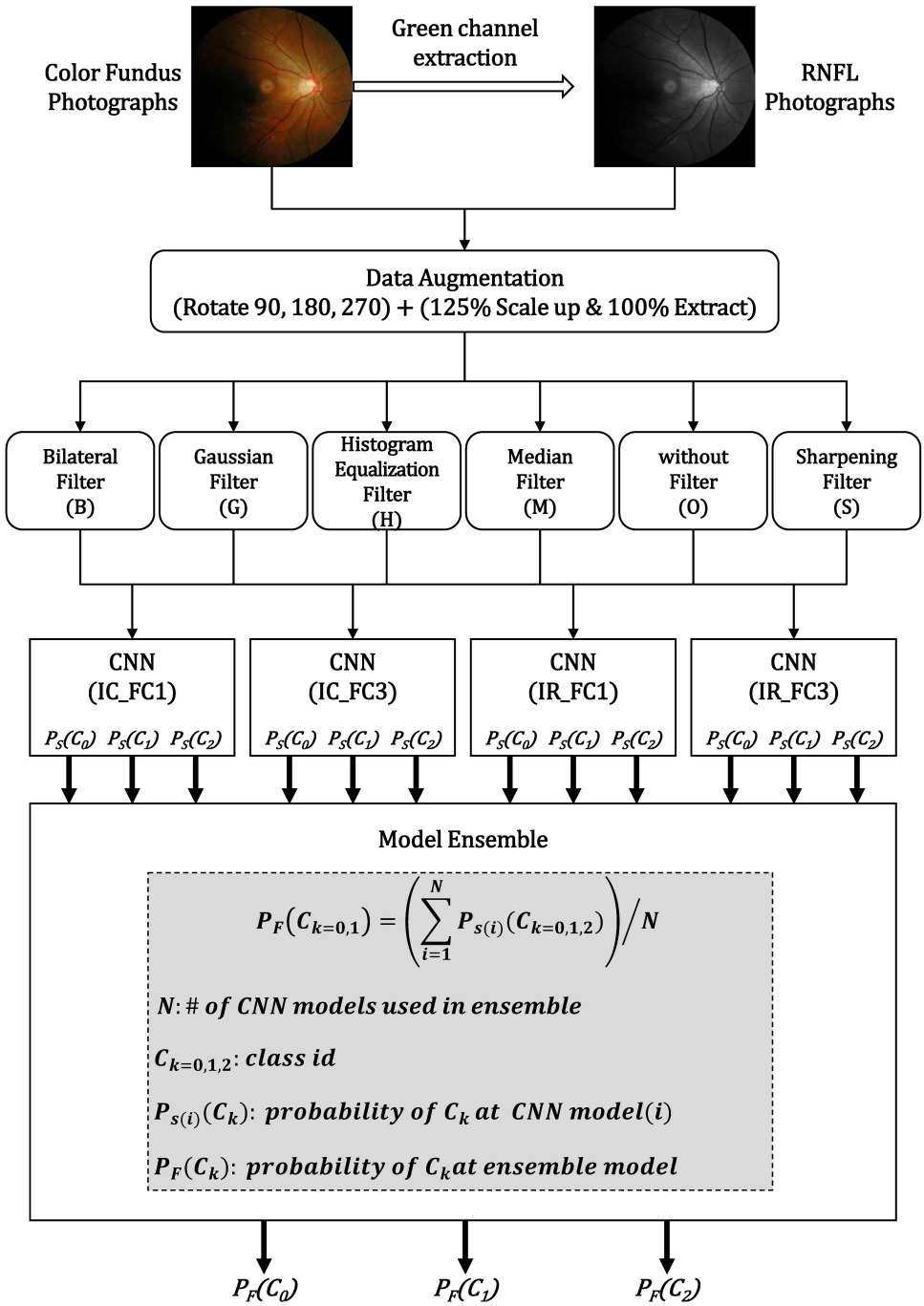


그림 22. 녹내장 중증도 등급화 앙상블 개념도

48개의 개별 CNN 모델을 훈련하였고, 학습이 완료된 모델을 대상으로 검증 데이터를 사용해서 성능을 평가하고 그 결과를 비교하였다. 학습이 완료된 48개의 개별 모델은 본 연구에서 제안하는 앙상블 방법에 따라 최종 판독 결과를 도출하였으며, 기준 모델과 앙상블 방법의 성능을 비교하였다. 그림21과 그림22는 앙상블 방법을 개념적으로 도식화하여 표현한 것이다.

본 연구에서 제안한 앙상블 방법은 개별 모델의 모든 출력을 합산하여 전체 개별 모델 수로 나누고, 그 결과가 가장 큰 값의 노드를 최종 결과로 결정하는 방법을 사용하였다. 예를 들어, 3개의 개별 모델이 있고, 각 모델에서 출력한 정상일 확률이 33%, 55%, 65%라면, 최종 결과는 (33%+55%+65%)/3을 계산하여 51%가 되고, 녹내장 확률이 67%, 45%, 35%라면, (67%+45%+35%)/3을 계산하여 49%가 돼서 최종적으로 정상으로 판독하는 것이다.

제 5 절 기계학습 실험 환경

본 연구에서 딥러닝 기계학습을 위해 사용한 하드웨어 및 소프트웨어 환경은 표5와 같다.

기계학습을 위한 기반 프레임웍은 텐서플로(TensorFlow) 1.8을 사용하였고, 개발 언어는 파이선 3.5를 사용하였다. 데이터 전처리를 위한 영상처리 도구는 OpenCV 3.1을 사용하였으며, 합성곱신경망의 고속 최적화를 위해 CUDA 9.0 및 5,120 cuda 코어를 지원하는 GPU를 사용하였다.

표 5. 딥러닝 기계학습 실험 환경

	규격/환경	비고
CPU	Intel Xeon Gold 6154	24.75M Cache, 3.00 GHz, 18 core
Main Memory	256GB	32GB ECC DDR4 SDRAM (2666Mhz) x 8 EA
Storage	SSD 6TB	SAMSUNG 860 Pro 2TB x 3 EA
GPU	Nvidia Titan V	1200 MHz, 12GB RAM, 5120 Cuda Cores
OS	Linux (Ubuntu 16.04x64)	-
Language	Python 3.6x64	-
OpenCV	openCV 3.1	-
Deep learning Framwork	Tensorflow 1.8	-
CUDA	cuda 9.0 & cudnn 7.0	-

딥러닝 학습 과정에서 사용한 Tensorflow 프레임웍의 학습 파라미터는 표6와 같이 설정하였다.

표 6. 딥러닝 기계학습 파라미터

파라미터	설정값
Batch size	64
Max epoch	1,500
Max fine tune	150
Channel size	3
Dropout	0.5
Learning rate	0.0001
Activation function	Relu
Optimizer	Adam
Loss	Categorical crossentropy

제 6 절 모델 평가 방법

딥러닝 학습이 완료된 결과물은 기계학습의 10-fold 교차검증 방법론에 따라 정확도, 민감도, 특이도, AUROC 지표를 기준으로 성능을 평가하였다. 성능평가를 위해 기계학습 데이터 구축 결과 중에서 10%를 별도로 임의 추출하여 테스트 데이터 집합으로 정의하고, 성능평가에 사용하였다. 또한, 테스트 데이터 집합은 기계학습 과정에는 포함하지 않은 데이터만으로 클래스별로 모두가 균등하게 분포하도록 구성하였다.

제 3 장 연구 결과

제 1 절 안저영상 DB 구축 결과

본 연구에서 합성곱신경망의 학습을 위해 구축한 전체 안저영상은 2,801명의 환자로부터 측정한 4,445장의 고해상도 컬러 영상으로 구성하였다. 구축한 영상의 전체 영상에 대해서 녹내장 전문의의 교차검증을 통해 최종적으로 딥러닝 기계학습을 위해 선별한 안저영상은 3,460이며, 환자 수는 2,204명이다.

교차검증을 시행하기 전 녹내장 등급별 분포는 정상이 1,848장으로 42%를 차지하고, 시야결손전녹내장(전기 녹내장)은 284장으로 6%이며, 초기 녹내장은 1,045장으로 23%이고, 중기 녹내장은 570장으로 13%이며, 말기 녹내장은 698장으로 16%의 분포를 이루고 있다.

교차검증을 거친 녹내장 등급별 분포는 정상 1,259장으로 37%를 차지하고, 시야결손전녹내장(전기 녹내장)은 185장으로 5%이며, 초기 녹내장은 784장으로 23%이고, 중기 녹내장은 563장으로 16%이며, 말기 녹내장은 669장으로 19%의 분포를 이루고 있다.

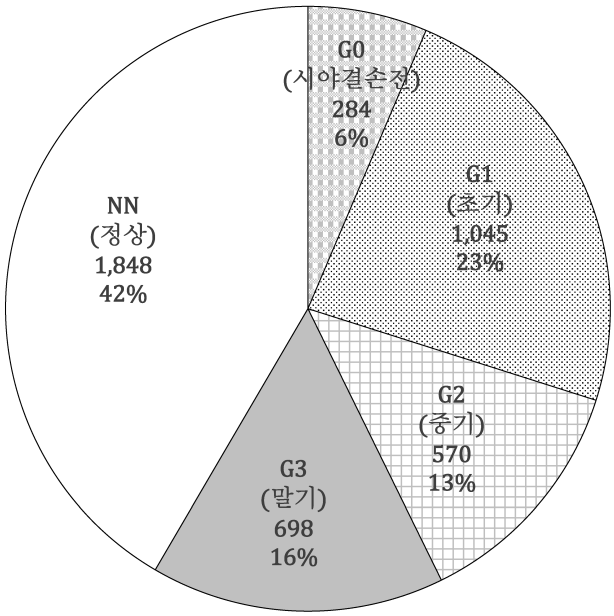
안저영상의 교차검증에 있어서 사진의 품질이 열악하거나 녹내장 전문의의 판독 결과가 일치하지 않는 안저영상은 모두 제외하였다. 교차검증을 통해 제외한 안저영상의 수량은 985장이다.

제외된 안저영상 중에 녹내장 등급의 불일치가 발생한 비율은 대략 50% 수준이다. 녹내장 전문의 사이에서 판독에 불일치가 발생한 내용을 검토해보면, 대부분 정상과 초기 녹내장 사이에 있는 영상들이었으며, 불일치 정도는 모두 1등급 이내였다.

표 7. 안저영상 구축 결과 및 교차검증 전후 비교

Grade (Code)	Data cross validation		Difference
	Before	After	
Unaffected Control (NN)	1,848	1,259	589
Preperimetric Glaucoma Grade (G0)	284	185	99
Mild Glaucoma Grade (G1)	1,045	784	261
Moderate Glaucoma Grade (G2)	570	563	7
Severe Glaucoma Grade (G3)	698	669	29
Total images	4,445	3,460	985
# of patients	2,801	2,204	597

(a). 데이터 교차 검증 전 분포



(b). 데이터 교차 검증 후 분포

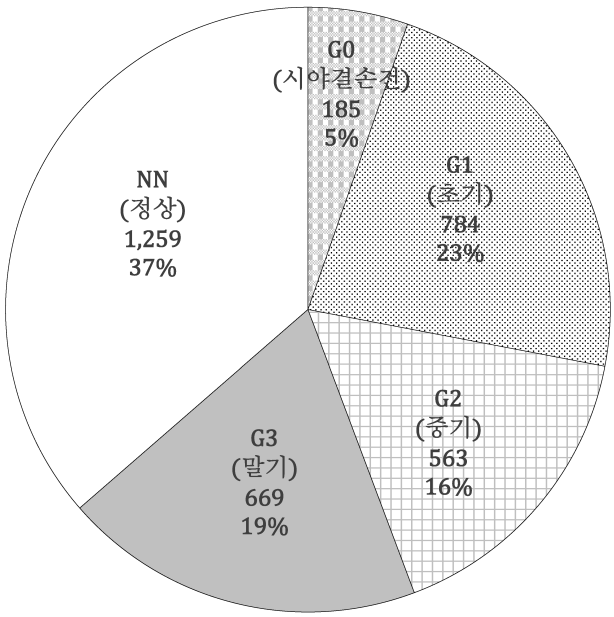


그림 23. 데이터 교차검증 전/후 분포

제 2 절 딥러닝 학습 및 테스트 데이터 구성

본 연구에서 기계학습을 위해 구성한 학습 데이터와 테스트 데이터의 규모는 표8과 같다. 녹내장 선별검사를 위한 이진 분류에서는 데이터 증폭이 이루어진 전체 데이터 집합에서 무작위 추출을 통해 총 9,792장의 안저영상을 학습에 사용하였다. 기계학습 데이터의 분포는 정상과 녹내장이 균등하게 분포하도록 무작위 추출하여 구성하였다. 또한, 녹내장 클래스의 중증도 구성 비율 역시 균등하게 분포하도록 하였다.

녹내장 선별검사를 위한 이진 분류 모델의 테스트 데이터 역시 데이터 증폭이 이루어진 전체 데이터 집합에서 학습 데이터 집합과 중복이 발생하지 않도록 무작위 추출하여 1,088장의 안저영상을 테스트에 사용하였다. 테스트 데이터의 분포 역시 기계학습 데이터와 마찬가지로 정상과 녹내장이 균등하게 분포하도록 무작위 추출하여 구성하였다. 녹내장 클래스 역시 중증도의 구성 비율이 균등하게 분포하도록 하였다.

녹내장 중등도 등급화를 위한 삼진 분류에서는 데이터 증폭이 이루어진 전체 데이터에서 무작위 추출로 7,344장의 안저영상을 훈련 집합으로 사용하였고, 이진 분류와 같은 방법으로 무작위 추출하여 학습 데이터의 클래스가 균등하게 분포하도록 구성하였다.

녹내장 중등도 등급화를 위한 삼진 분류에서의 테스트 데이터 역시 데이터 증폭이 이루어진 전체 데이터에서 무작위로 816장의 안저영상을 테스트에 사용하였다. 테스트 데이터의 분포 역시 이진 분류 기계학습에서와 마찬가지로 녹내장 중증도 별로 균등하게 분포하도록 무작위 추출하여 구성하였다.

표 8. 학습 및 테스트 데이터 구성

모델 구분	클래스 (코드)	등급 코드	학습 데이터	비율	테스트 데이터	비율
녹내장 선별검사 (이진 분류)	Unaffected Control =정상 (C0)	NN	4,896	50.0%	544	50.0%
	Glaucoma =녹내장 (C1)	G0	1,224	12.5%	136	12.5%
		G1	1,224	12.5%	136	12.5%
		G2	1,224	12.5%	136	12.5%
		G3	1,224	12.5%	136	12.5%
		소계	4,896	50.0%	544	50.0%
	합계		9,792	100.0%	1,088	100.0%
녹내장 중증도 등급화 (삼진 분류)	Unaffected Control =정상 (C0)	NN	2,448	33.3%	272	33.3%
	Early-Stage Glaucoma= 초기 녹내장 (C1)	G0	1,224	16.7%	136	16.7%
		G1	1,224	16.7%	136	16.7%
		소계	2,448	33.3%	272	33.3%
	Late-Stage Glaucoma= 중증 녹내장 (C2)	G2	1,224	16.7%	136	16.7%
		G3	1,224	16.7%	136	16.7%
		소계	2,448	33.3%	272	33.3%
	합계		7,344	100.0%	816	100.0%

제 3 절 녹내장 선별검사 모델 학습결과

1. 녹내장 선별검사 개별 모델 학습결과

본 연구에서는 녹내장 선별검사를 위한 이진 분류 모델의 성능 비교를 위해 InceptionNet-v3 모델에 어떠한 영상 필터도 적용하지 않은 방법을 기준 모델로 사용하였다. 학습에 사용한 데이터 증폭 방법은 Li et al. (2018)에서 제시한 방법을 적용해서 본 연구에서 제시한 앙상블 방법과 비교였다.

Li et al. (2018)에서 제시한 방법과 본 연구에서 제안하는 앙상블 방법을 비교하기 위해서 InceptionNet-v3 모델과 Inception-ResNet-v2 모델의 일부 계층을 그림19와 같이 수정하여 표3과 같이 48개의 개별 모델을 훈련시켰다. 개별 모델 구성 방법으로는 원본 입력 데이터의 패턴을 실험 설계에서 정의한 바와 같이 영상확대, 영상 회전 방법으로 데이터를 증폭하였고, 안저 영상의 컬러 형태별로 표2에서 제시한 영상 필터를 적용하여 48개의 개별 모델을 구성하고 기계학습을 진행하였다.

개별 모델의 훈련 과정을 모델 정확도를 기준으로 살펴보면, InceptionNet-v3는 400회 반복까지는 학습률이 급격하게 증가하였고, Inception-ResNet-v2는 200회 반복까지 학습률이 급격하게 증가하였다. 반복 횟수를 기준으로 보면 Inception-ResNet-v2가 빠르게 학습하는 것처럼 보인다. 그러나, Inception-ResNet-v2의 1회 반복 시간이 InceptionNet-v3보다 3배 이상 소요된다. 따라서, InceptionNet-v3이 최종적인 학습 완료까지 빠르게 도달함을 확인하였고, GPU 메모리 사용 효율 측면에서도 유리하게 작용하는 것으로 나타났다.

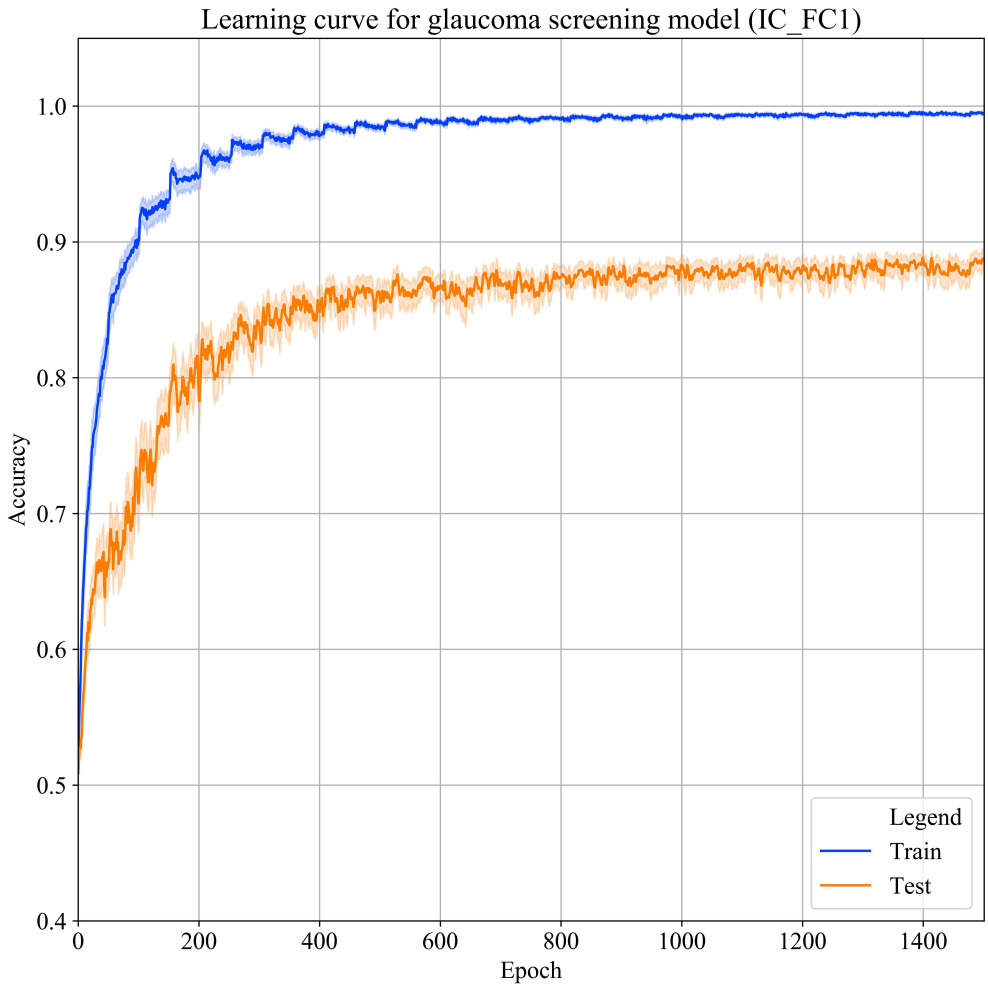


그림 24. 녹내장 선별검사 IC_FC1 학습 곡선

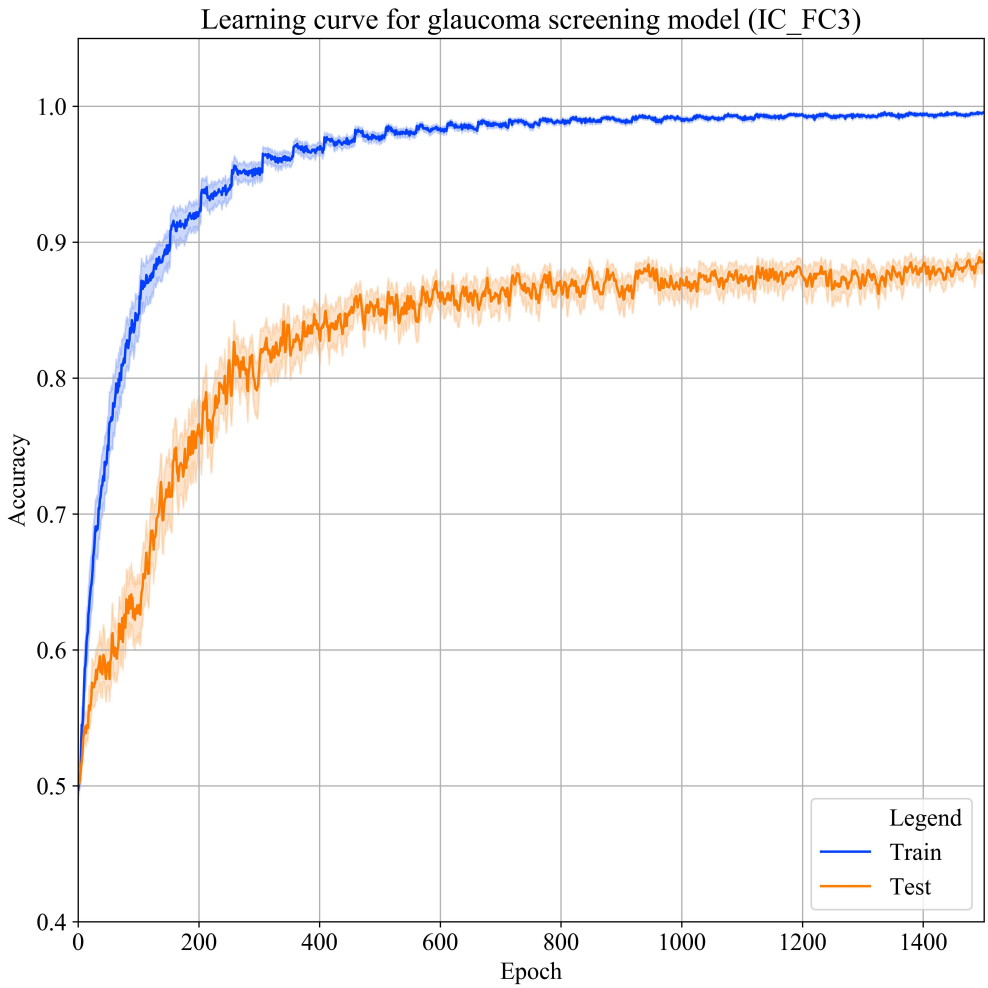


그림 25. 녹내장 선별검사 IC_FC3 학습 곡선

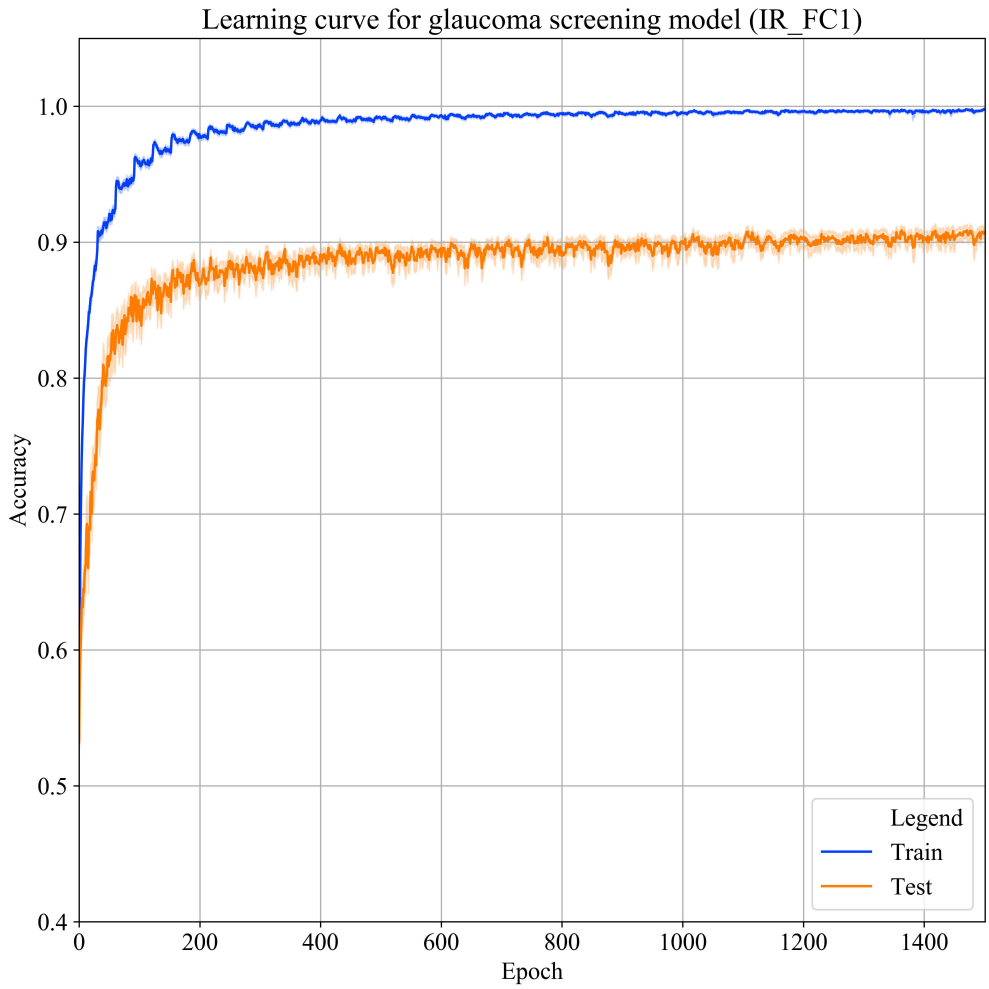


그림 26. 녹내장 선별검사 IR_FC1 학습 곡선

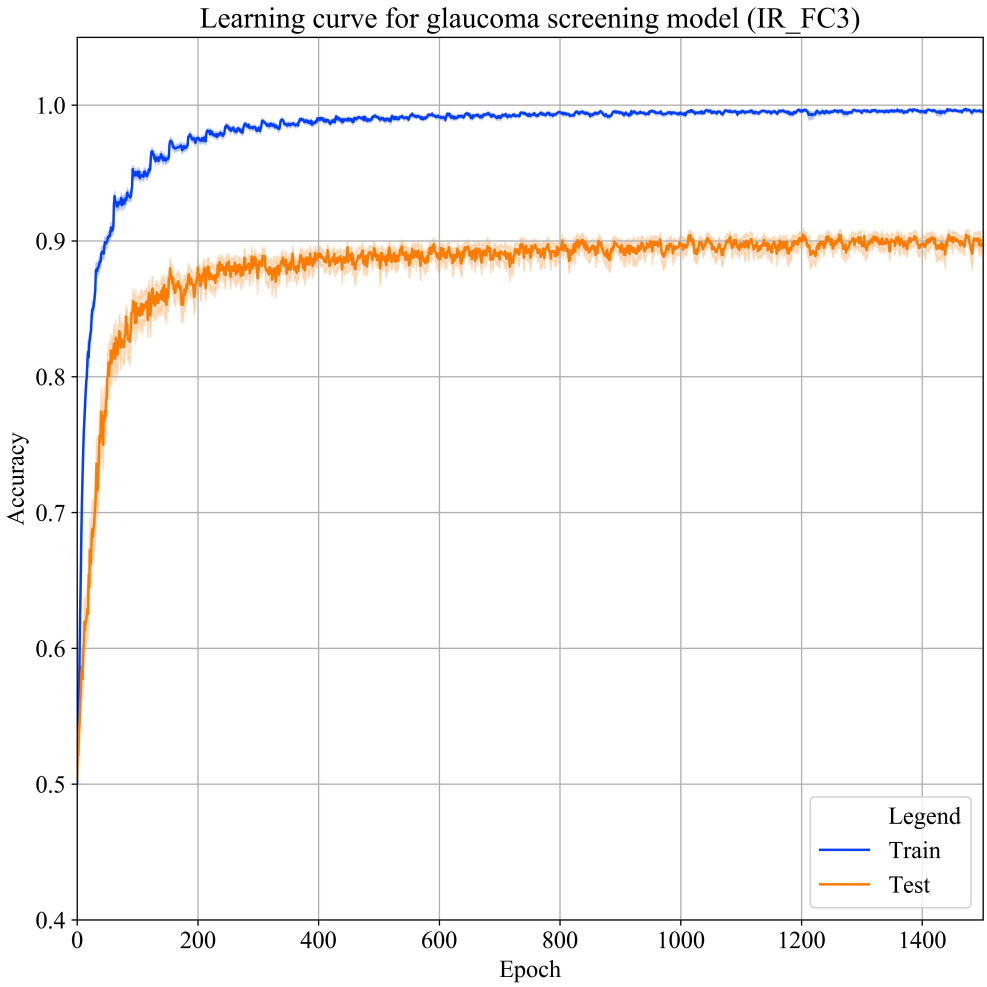


그림 27. 녹내장 선별검사 IR_FC3 학습 곡선

녹내장 선별검사를 위한 이진 분류 개별 모델 중에서 AUROC 기준으로 가장 좋은 성능을 보이는 모델은 CNN(2C)_G_G_IR_FC1로 평균 0.981 (95% confidence interval [CI], 0.974 ~ 0.988, SD=0.0097)을 보였고, 정확도를 기준으로 가장 좋은 성능을 보이는 모델은 CNN(2C)_G_C_IR_FC3로 평균 94.6% (95% CI, 93.2 ~ 95.9%, SD=0.0174)를 보였으며, 민감도 기준으로 가장 좋은 성능을 보이는 모델은 CNN(2C)_G_C_IC_FC3로 평균 95.0% (95% CI, 93.3 ~ 96.7%, S=0.0230)을 보였고, 특이도 기준으로 가장 좋은 성능을 보이는 모델은 CNN(2C)_S_C_IR_FC3로 평균 95.9% (95% CI, 94.2 ~ 97.6%, SD=0.0226)의 성능을 보였다.

표 9. 녹내장 선별검사 개별 모델에서 지표별 최고 성능

성능 지표	최고 성능 개별 모델	성능	신뢰구간	표준편차
AUROC	CNN(2C)_G_G_IR_FC1	0.981	(95% CI, 0.974 ~ 0.988)	0.0097
Accuracy	CNN(2C)_G_C_IR_FC3	94.6%	(95% CI, 93.2 ~ 95.9%)	0.0174
Sensitivity	CNN(2C)_G_C_IC_FC3	95.0%	(95% CI, 93.3 ~ 96.7%)	0.0230
Specificity	CNN(2C)_S_C_IR_FC3	95.9%	(95% CI, 94.2 ~ 97.6%)	0.0226

2. 녹내장 선별검사 앙상블 학습결과

학습한 48개의 녹내장 선별검사 개별 모델을 본 연구에서 제안한 앙상블 방법을 적용한 판독 결과를 살펴보면, 아래 그림과 표에서 보는 바와 같이 모든 지표에서 우월함을 확인할 수 있다.

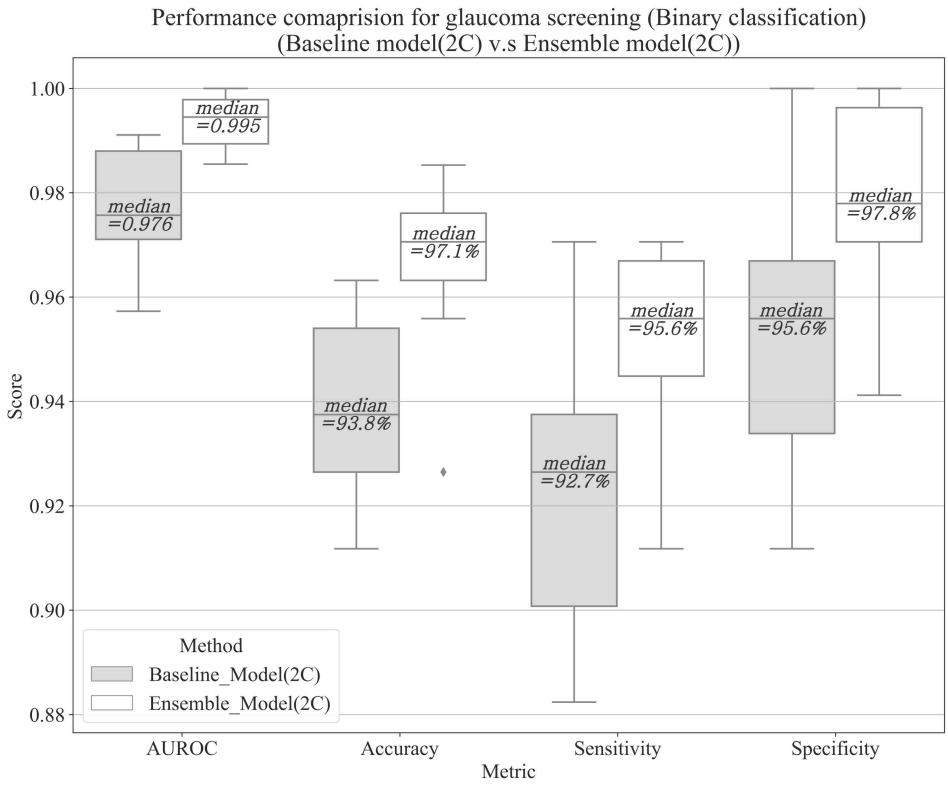


그림 28. 녹내장 선별검사 모델 성능 비교

녹내장 선별검사의 정확도 측면에서 앙상블 방법은 96.6% (95% CI, 95.5 ~ 97.8%)를 보였다. 반면, InceptionNet-v3 모델 한 개를 사용한 기준 모델은 93.9% (95% CI, 92.6 ~ 95.2%)를 보였다. 기준 모델과 앙상블 방법의 녹내장 선별검사 정확도에 대한 성능 차이를 paired t-test로 통계적 유의성을 검정하였으며, 결과는 p-value 0.000425로 정확도의 차이가 통계적으로 유의함을 밝혔다.

AUROC 측면에서 앙상블 방법은 0.994 (95% CI, 0.990 ~ 0.997)를 보였으며, InceptionNet-v3 모델 한 개를 사용한 기준 모델은 0.977 (95% CI, 0.969 ~ 0.986)를 보였다. 녹내장 선별검사에 있어서 기준 모델과 앙상블 방법의 AUROC에 대한 성능 차이를 paired t-test로 통계적 유의성을 검정하였으며, 결과는 p-value 0.000966으로 AUROC의 차이가 통계적으로 유의함을 밝혔다.

민감도 측면에서 앙상블 방법은 95.3% (95% CI, 94.0 ~ 96.6%)를 보였으며, 기준 모델은 92.4% (95% CI, 90.4 ~ 94.3%)를 보였다. 민감도의 paired t-test 검정 결과 p-value는 0.014956였다.

특이도 측면에서 앙상블 방법은 97.9% (95% CI, 96.6 ~ 99.3%)를 보였으며, 기준 모델은 95.4% (95% CI, 93.5 ~ 97.4%)를 보였다. 특이도의 paired t-test 검정 결과 p-value는 0.002005였다.

이로써 녹내장 선별검사에서 앙상블 방법이 정확도와 AUROC 측면에서 더 높고 안정적인 것을 확인하였다. 또한, 민감도와 특이도의 관점에서도 본 연구에서 제안하는 앙상블 모델이 모든 지표의 평균, 최대, 최소, 분산을 비교해 볼 때, 월등히 우수함을 확인할 수 있었다.

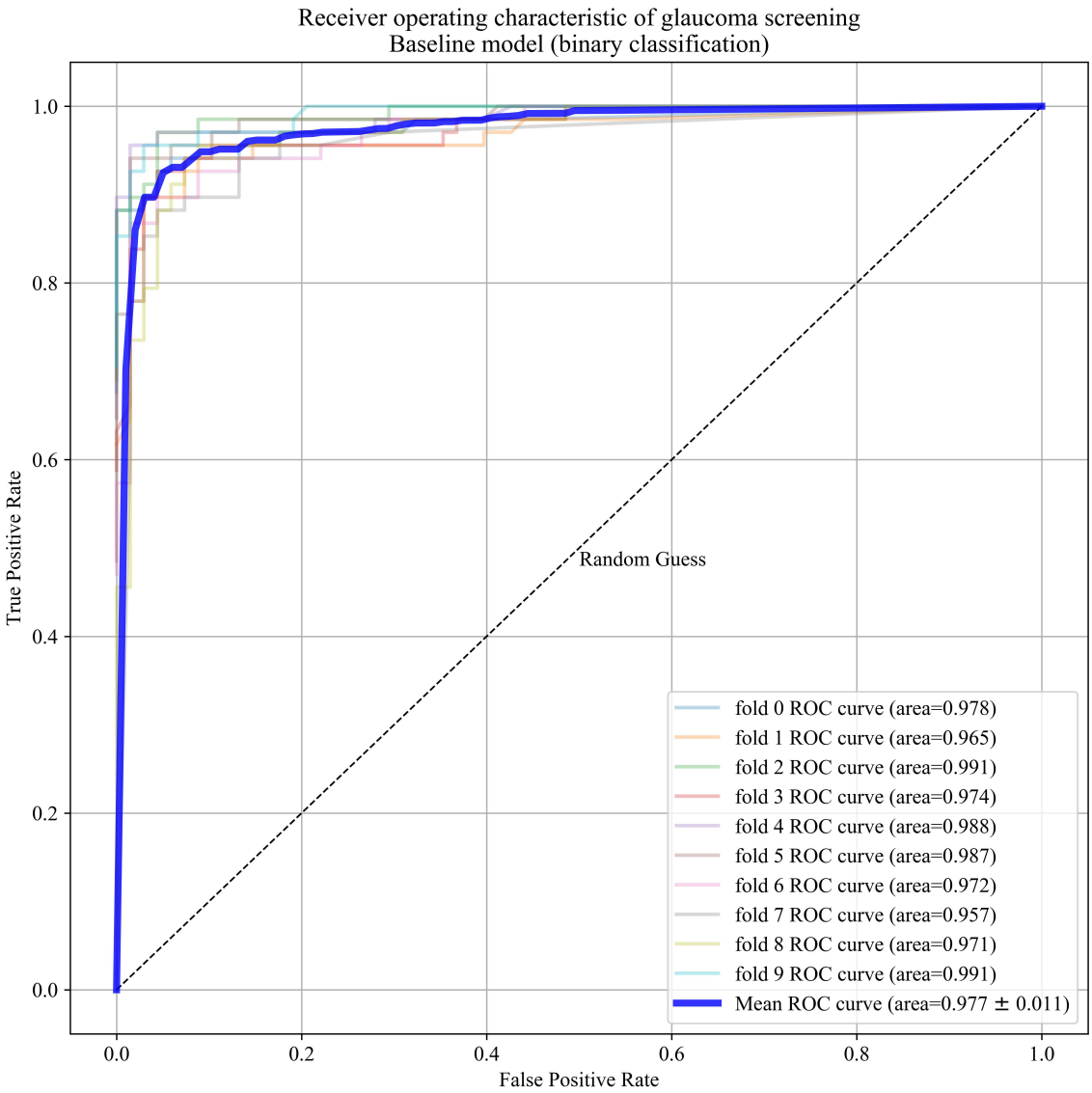


그림 29. 녹내장 선별검사 기준 모델 ROC 곡선

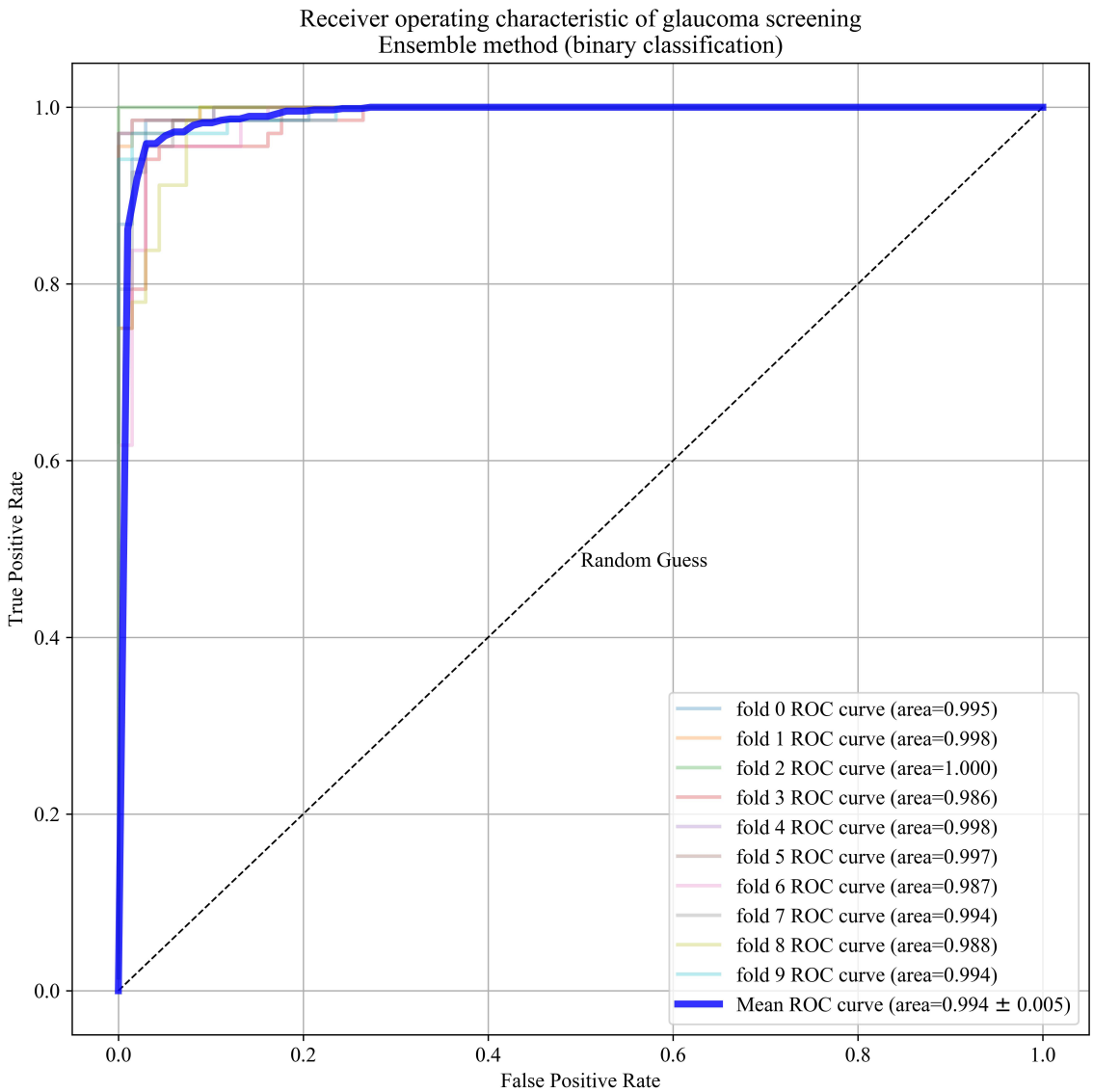


그림 30. 녹내장 선별검사 앙상블 방법 ROC 곡선

표 10. 녹내장 선별검사 모델 성능

	AUROC		Accuracy		Sensitivity		Specificity	
	Baseline model (2C)	Ensemble model (2C)	Baseline model (2C)	Ensemble model (2C)	Baseline model (2C)	Ensemble model (2C)	Baseline model (2C)	Ensemble model (2C)
fold0	0.978	0.995	94.1%	97.1%	92.7%	97.1%	95.6%	97.1%
fold1	0.965	0.998	93.4%	97.8%	89.7%	97.1%	97.1%	98.5%
fold2	0.991	1.000	96.3%	98.5%	97.1%	97.1%	95.6%	100.0%
fold3	0.974	0.986	92.7%	95.6%	92.7%	94.1%	92.7%	97.1%
fold4	0.988	0.998	94.9%	97.1%	89.7%	94.1%	100.0%	100.0%
fold5	0.987	0.997	96.3%	97.8%	94.1%	95.6%	98.5%	100.0%
fold6	0.972	0.987	91.2%	96.3%	91.2%	95.6%	91.2%	97.1%
fold7	0.957	0.994	91.9%	96.3%	88.2%	95.6%	95.6%	97.1%
fold8	0.971	0.988	92.7%	92.7%	92.7%	91.2%	92.7%	94.1%
fold9	0.991	0.994	95.6%	97.1%	95.6%	95.6%	95.6%	98.5%
평균	0.977	0.994	93.9%	96.6%	92.4%	95.3%	95.4%	97.9%
최소	0.957	0.986	91.2%	92.7%	88.2%	91.2%	91.2%	94.1%
최대	0.991	1.000	96.3%	98.5%	97.1%	97.1%	100.0%	100.0%

표 11. 녹내장 선별검사 성능 통계분석 결과

Metric	Group	Score	SD	Shapiro-Wilk normality test	Paired t-test
AUROC	Baseline model (2C)	0.977 (95% CI, 0.969 ~ 0.986)	0.0111	p-value= 0.069788	p-value= 0.000966
	Ensemble model (2C)	0.994 (95% CI, 0.990 ~ 0.997)	0.0049		
Accuracy	Baseline model (2C)	93.9% (95% CI, 92.6 ~ 95.2%)	0.0174	p-value= 0.767156	p-value= 0.000425
	Ensemble model (2C)	96.6% (95% CI, 95.5 ~ 97.8%)	0.0155		
Sensitivity	Baseline model (2C)	92.4% (95% CI, 90.4 ~ 94.3%)	0.0261	p-value= 0.331520	p-value= 0.014956
	Ensemble model (2C)	95.3% (95% CI, 94.0 ~ 96.6%)	0.0171		
Specificity	Baseline model (2C)	95.4% (95% CI, 93.5 ~ 97.4%)	0.0258	p-value= 0.102475	p-value= 0.002005
	Ensemble model (2C)	97.9% (95% CI, 96.6 ~ 99.3%)	0.0176		

❍ CI = Confidence interval

❍ SD = Standard Deviation

❍ AUROC = Area Under the Response Operating Characteristic curve

제 4 절 녹내장 중증도 등급화 모델 학습결과

1. 녹내장 중증도 등급화 개별 모델 학습결과

현재까지 딥러닝을 사용해 녹내장의 중증도를 등급화한 사례가 없으므로, 본 연구에서는 원본 영상에 어떠한 영상 필터도 적용하지 않은 입력을 사용한 InceptionNet-v3을 중증도 등급화에 맞게 변형하여 기준 모델로 사용하고, 기준 모델을 앙상블 방법과 비교하였다.

녹내장 중증도 등급화를 위해서 InceptionNet-v3 모델과 Inception-ResNet-v2 모델의 일부 계층을 그림20과 같이 수정하여 표4와 같이 48개의 개별 모델을 훈련시켰다. 개별 모델 구성 방법으로는 원본 입력 데이터의 패턴을 실험 설계에서 정의한 바와 같이 영상확대, 영상 회전 방법으로 데이터를 증폭하였고, 안저 영상의 컬러 형태별로 표2에서 제시한 영상 필터를 적용하여 48개의 개별 모델을 구성하고 기계학습을 진행하였다.

개별 모델의 훈련 과정을 모델 정확도를 기준으로 살펴보면, InceptionNet-v3는 600회 반복까지 학습률이 급격하게 증가하였고, Inception-ResNet-v2는 400회 반복까지 학습률이 급격하게 증가하였다.

중증도 등급화를 위한 삼진 분류 모델의 학습에서는 학습 데이터의 정확도가 100%까지 이르지 못하고, 개별 모델의 테스트 데이터 정확도 역시 대략 70% 근처에 머무르는 것을 확인할 수 있었다. 이러한 결과는 이진 분류에서 삼진 분류 문제로 확대되는 과정에서 더 많은 학습 데이터가 필요하다는 간접적 증거로 해석하였다.

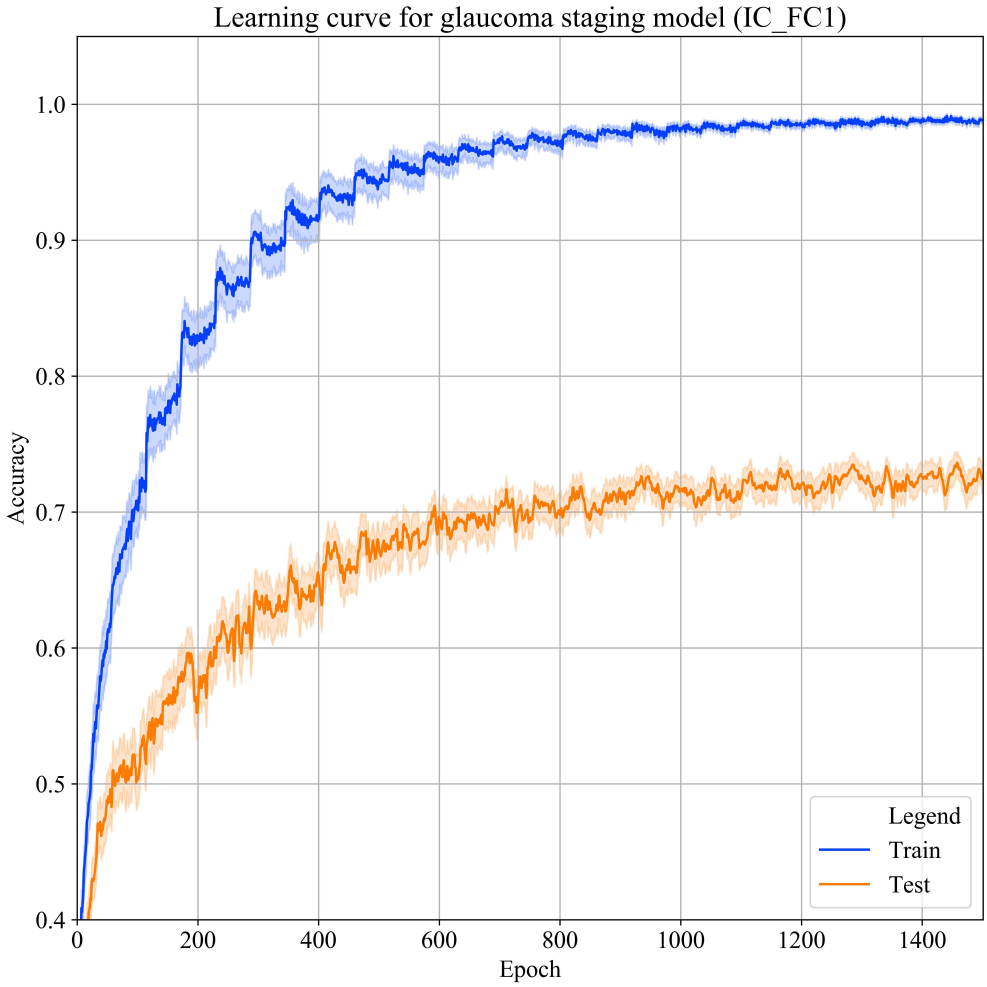


그림 31. 녹내장 중증도 등급화 IC_FC1 모델 학습 곡선

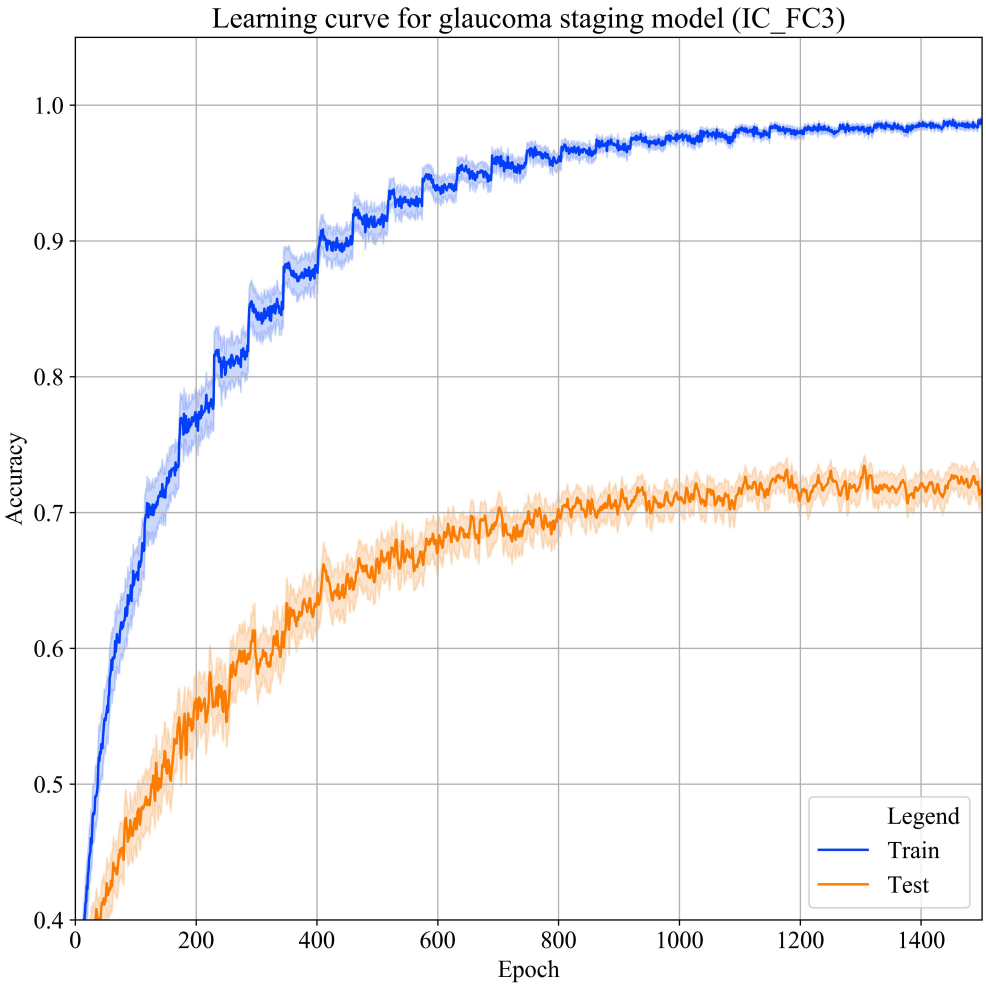


그림 32. 녹내장 중증도 등급화 IC_FC3 모델 학습 곡선

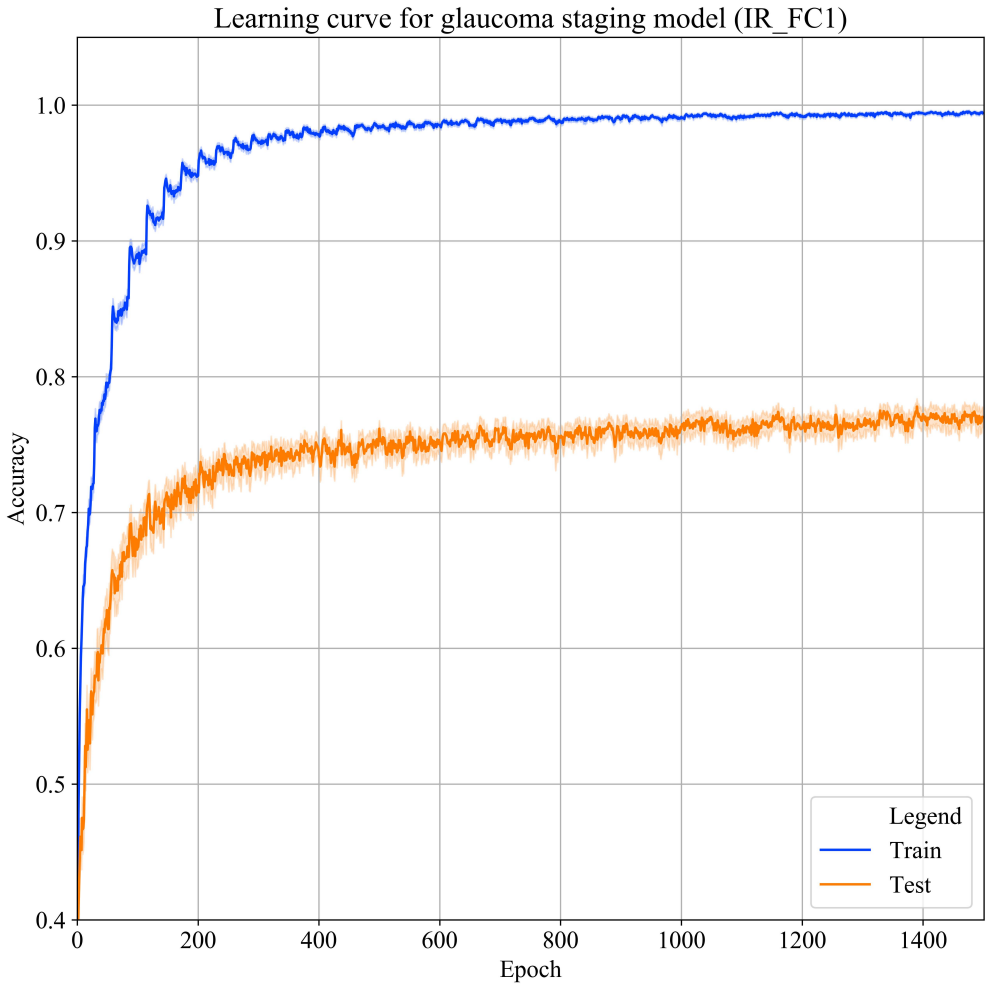


그림 33. 녹내장 중증도 등급화 IR_FC1 모델 학습 곡선

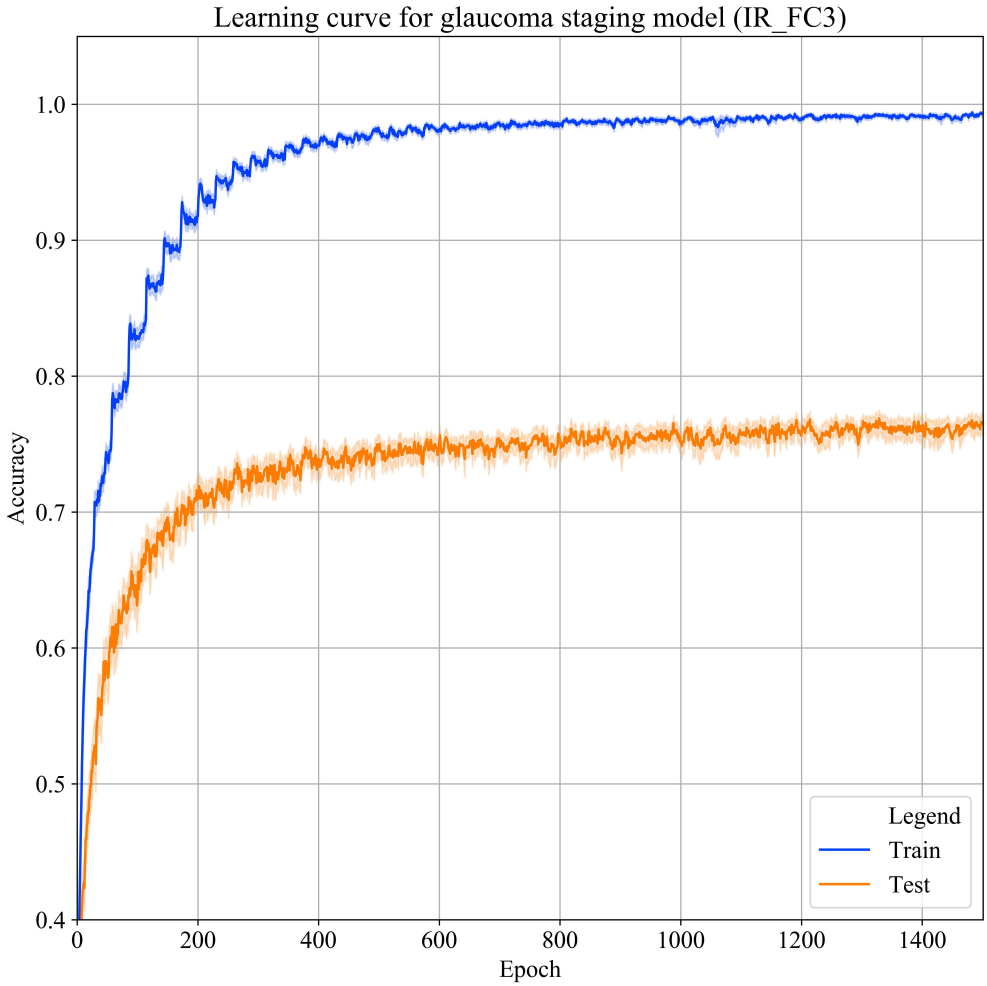


그림 34. 녹내장 중증도 등급화 IR_FC3 모델 학습 곡선

녹내장 중증도 등급화를 위한 개별 모델에서 성능 지표별로 가장 우수한 결과를 보이는 모델은 표12에 정리하였다. 녹내장 중증도 등급화를 위한 개별 모델 중에서 정확도 기준으로 가장 좋은 성능을 보이는 모델은 CNN(3C)_S_C_IR_FC1로 평균 85.2% (95% CI, 83.5 ~ 86.9%, SD=0.0230)를 보였고, Unaffected Control AUROC 기준으로 가장 좋은 성능을 보이는 모델은 CNN(3C)_S_C_IR_FC1로 평균 0.980 (95% CI, 0.972 ~ 0.987, SD=0.0096)을 보였으며, Early-Stage Glaucoma AUROC 기준으로 가장 좋은 성능을 보이는 모델은 CNN(3C)_M_C_IR_FC1로 평균 0.903 (95% CI, 0.881 ~ 0.925, S=0.0292)을 보였고, Late-Stage Glaucoma AUROC 기준으로 가장 좋은 성능을 보이는 모델은 CNN(3C)_O_C_IC_FC1로 평균 0.953 (95% CI, 0.936 ~ 0.970, SD=0.0222)의 성능을 보였다.

표 12. 녹내장 중증도 등급화 개별 모델에서 지표별 최고 성능

성능 지표	최고 개별 모델	성능	신뢰구간	표준편차
Accuracy	CNN(3C)_S_C_IR_FC1	85.2%	(95% CI, 83.5 ~ 86.9%)	0.0230
Unaffected Control AUROC	CNN(3C)_S_C_IR_FC1	0.980	(95% CI, 0.972 ~ 0.987)	0.0096
Early-Stage Glaucoma AUROC	CNN(3C)_M_C_IR_FC1	0.903	(95% CI, 0.881 ~ 0.925)	0.0292
Late-Stage Glaucoma AUROC	CNN(3C)_O_C_IC_FC1	0.953	(95% CI, 0.936 ~ 0.970)	0.0222

2. 녹내장 중증도 등급화 앙상블 학습결과

녹내장 중증도 등급화를 위한 48개의 개별 모델을 본 연구에서 제안한 앙상블 방법을 적용한 판독 성능은 아래 그림과 표에서 제시한 바와 같이 모든 지표에서 우월함을 확인할 수 있었다.

녹내장 중증도 등급화의 정확도 측면에서 앙상블 방법은 87.7% (95% CI, 85.9 ~ 89.7%)를 보였다. 반면, InceptionNet-v3 모델 한 개를 사용한 기준 모델은 82.3% (95% CI, 80.2 ~ 84.1%)를 보였다. 기준 모델과 앙상블 방법의 녹내장 중증도 등급화 정확도에 대한 성능 차이를 paired t-test로 통계적 유의성을 검정하였고, 결과는 p-value 0.002902로 정확도의 차이가 통계적으로 유의함을 밝혔다.

평균 AUROC 측면에서 앙상블 방법은 0.975 (95% CI, 0.967 ~ 0.983)를 보였으며, InceptionNet-v3 모델 한 개를 사용한 기준 모델은 0.938 (95% CI, 0.926 ~ 0.949)을 보였다. 평균 AUROC 측면에서 기준 모델과 앙상블 방법의 성능 차이를 paired t-test로 통계적 유의성을 검정하였고, 결과는 p-value 0.000093으로 평균 AUROC의 차이가 통계적으로 유의함을 밝혔다.

정상을 식별하는 AUROC 측면에서 앙상블 방법은 0.990 (95% CI, 0.987 ~ 0.994)를 보였으며, InceptionNet-v3 모델 한 개를 사용한 기준 모델은 0.962 (95% CI, 0.952 ~ 0.974)를 보였다. 정상을 식별하는 AUROC 측면에서 기준 모델과 앙상블 방법의 성능 차이를 paired t-test로 통계적 유의성을 검정하였고, 결과는 p-value 0.000496으로 정상을 식별하는 AUROC의 차이가 통계적으로 유의함을 밝혔다.

초기 녹내장을 식별하는 AUROC 측면에서 앙상블 방법은 0.951 (95% CI, 0.939 ~ 0.962)를 보였으며, InceptionNet-v3 모델 한 개를 사용한 기

준 모델은 0.882 (95% CI, 0.862 ~ 0.898)를 보였다. 초기 녹내장을 식별하는 AUROC 측면에서 기준 모델과 앙상블 방법의 성능 차이를 paired t-test로 통계적 유의성을 검정하였고, 결과는 p-value 0.00018로 초기 녹내장을 식별하는 AUROC의 차이가 통계적으로 유의함을 밝혔다.

중증 녹내장을 식별하는 AUROC 측면에서 앙상블 방법은 0.970 (95% CI, 0.958 ~ 0.981)을 보였으며, InceptionNet-v3 모델 한 개를 사용한 기준 모델은 0.953 (95% CI, 0.938 ~ 0.968)을 보였다. 중증 녹내장을 식별하는 AUROC 측면에서 기준 모델과 앙상블 방법의 성능 차이를 paired t-test로 통계적 유의성을 검정하였고, 결과는 p-value 0.016406으로 중증 녹내장을 식별하는 AUROC의 차이가 통계적으로 유의함을 밝혔다.

이로써 녹내장 중증도 등급화에서도 본 연구에서 제안하는 앙상블 방법이 정확도, AUROC의 평균, 최대, 최소, 분산을 비교해 볼 때, 더 높고 안정적이며, 월등히 우수함을 확인할 수 있었다.

녹내장 중증도 등급화 모델의 성능 비교 결과에서 특히 중요한 내용은 앙상블 방법의 초기 녹내장 진단 능력이다. 앙상블 방법의 초기 녹내장 진단 능력은 기준 모델과 비교할 때, 월등히 우수함을 확인할 수 있었다. 이러한 특징은 녹내장 질병의 관점에서 임상적으로 매우 큰 의미가 있다. 서론에서 설명한 바와 같이 초기 녹내장을 진단하는 능력은 환자의 치료 예후에 중대한 영향을 주며, 궁극적으로 환자의 실명 위험을 크게 낮출 수 있기 때문이다.

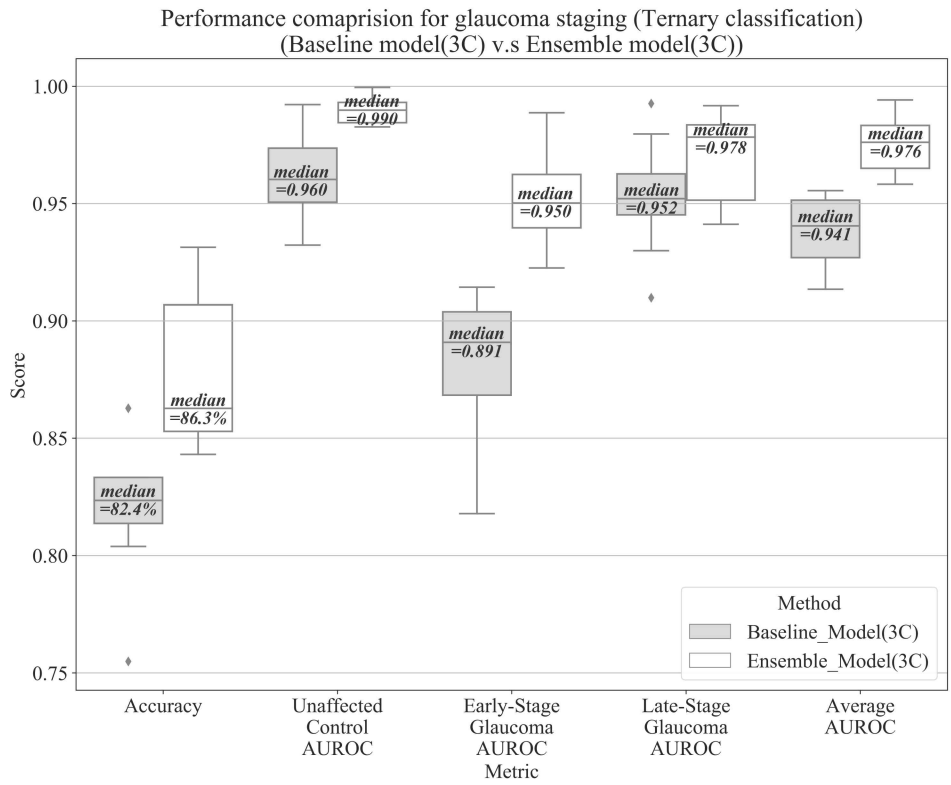


그림 35. 녹내장 중증도 등급화 모델 성능 비교

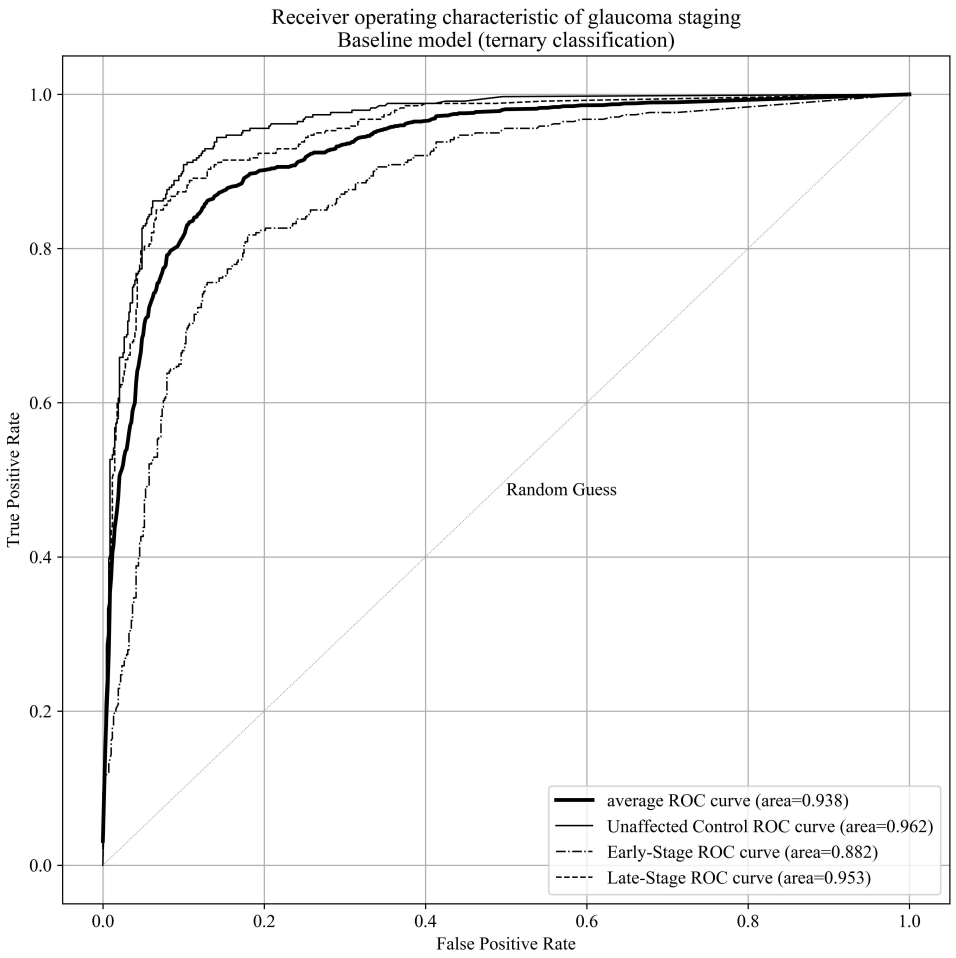


그림 36. 녹내장 중증도 등급화 기준 모델 ROC 곡선

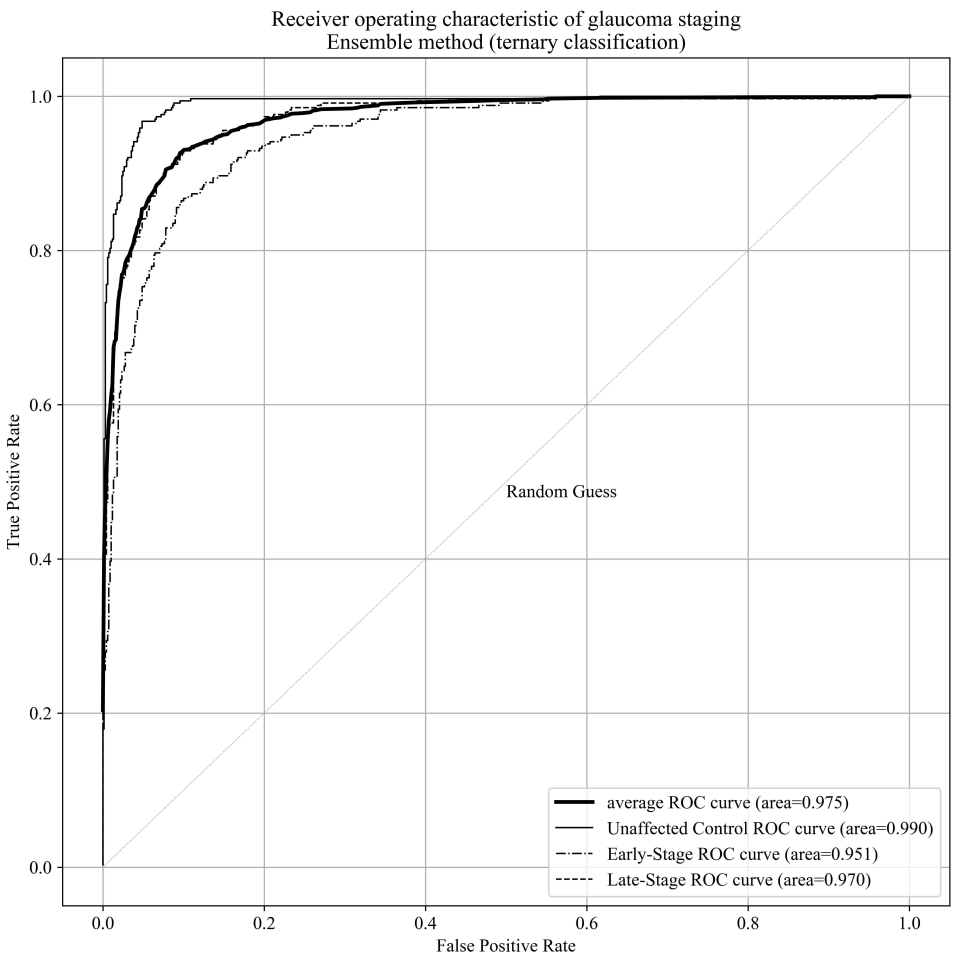


그림 37. 녹내장 중증도 등급화 앙상블 모델 ROC 곡선

표 13. 녹내장 중증도 등급화 성능 - 정확도 및 평균 AUROC

Fold	Accuracy		Average AUROC	
	Baseline model (3C)	Ensemble model (3C)	Baseline model (3C)	Ensemble model (3C)
fold0	81.4%	84.3%	0.929	0.969
fold1	83.3%	86.3%	0.934	0.958
fold2	81.4%	93.1%	0.953	0.994
fold3	86.3%	89.2%	0.948	0.978
fold4	83.3%	85.3%	0.956	0.977
fold5	80.4%	91.2%	0.917	0.985
fold6	83.3%	92.2%	0.947	0.986
fold7	86.3%	86.3%	0.953	0.962
fold8	81.4%	84.3%	0.927	0.964
fold9	75.5%	85.3%	0.914	0.976
평균	82.3%	87.7%	0.938	0.975
최소	75.5%	84.3%	0.914	0.958
최대	86.3%	93.1%	0.956	0.994

표 14. 녹내장 중증도 등급화 성능 - AUROC

Fold	Unaffected Control AUROC		Early-Stage Glaucoma AUROC		Late-Stage Glaucoma AUROC	
	Baseline model (3C)	Ensemble model (3C)	Baseline model (3C)	Ensemble model (3C)	Baseline model (3C)	Ensemble model (3C)
fold0	0.968	0.984	0.867	0.955	0.930	0.941
fold1	0.959	0.983	0.873	0.923	0.955	0.950
fold2	0.984	0.999	0.909	0.989	0.950	0.989
fold3	0.958	0.986	0.890	0.954	0.980	0.982
fold4	0.945	0.983	0.914	0.946	0.993	0.992
fold5	0.948	0.992	0.818	0.965	0.965	0.984
fold6	0.976	1.000	0.899	0.967	0.954	0.982
fold7	0.992	0.994	0.906	0.930	0.948	0.947
fold8	0.961	0.988	0.891	0.938	0.910	0.955
fold9	0.932	0.992	0.849	0.946	0.944	0.975
평균	0.962	0.990	0.882	0.951	0.953	0.970
최소	0.932	0.983	0.818	0.923	0.910	0.941
최대	0.992	1.000	0.914	0.989	0.993	0.992

표 15. 녹내장 중증도 등급화 성능 통계분석 결과

Metric	Group	Score	SD	Shapiro-Wilk normality test	Paired t-test
Accuracy	Baseline model (3C)	82.3% (95% CI, 80.2 ~ 84.1%)	0.0312	p-value = 0.053193	p-value = 0.002902
	Ensemble model (3C)	87.7% (95% CI, 85.9 ~ 89.7%)	0.0337		
Average AUROC	Baseline model (3C)	0.938 (95% CI, 0.926 ~ 0.949)	0.0148	p-value = 0.7199	p-value = 0.000093
	Ensemble model (3C)	0.975 (95% CI, 0.967 ~ 0.983)	0.011		
Unaffected Control AUROC	Baseline model (3C)	0.962 (95% CI, 0.952 ~ 0.974)	0.0183	p-value = 0.854167	p-value = 0.000496
	Ensemble model (3C)	0.990 (95% CI, 0.987 ~ 0.994)	0.0063		
Early-Stage Glaucoma AUROC	Baseline model (3C)	0.882 (95% CI, 0.862 ~ 0.898)	0.0304	p-value = 0.556333	p-value = 0.000180
	Ensemble model (3C)	0.951 (95% CI, 0.939 ~ 0.962)	0.0193		
Late-Stage Glaucoma AUROC	Baseline model (3C)	0.953 (95% CI, 0.938 ~ 0.968)	0.0234	p-value = 0.293418	p-value = 0.016406
	Ensemble model (3C)	0.970 (95% CI, 0.958 ~ 0.981)	0.0193		

❍ CI = Confidence interval

❍ SD = Standard Deviation

❍ AUROC = Area Under the Response Operating Characteristic curve

제 4 장 고 찰

제 1 절 딥러닝 기반 녹내장 판독 시스템 개발

본 연구에서는 안저영상만으로 초기 녹내장까지도 판독할 수 있는 합성곱신경망 모델을 제시하였다. 안저영상은 데이터를 확보하는 방법과 비용, 시간 측면에서 많은 장점이 있어서 널리 보급된 안과 의료기기 중의 하나이다. 특히, 본 연구에서 사용한 안저 영상은 산동 후 팽창 임상 안저촬영 결과를 사용한 것이 아니고, 보통 상태에서 촬영한 결과를 사용했다는 특징으로 인해 향후 적용방법의 보편성 측면에서 그 의미가 크다고 할 수 있다.

연구에 사용한 데이터의 또 다른 특징으로는 녹내장 전문의가 안저영상의 구조적 특성만을 가지고 영상에 녹내장 등급을 할당한 것이 아니라, 임상 현장에서 녹내장 확진에 가장 보편적으로 사용하는 기능 검사법인 시야검사 결과와 OCT 촬영 결과를 안저영상과 동시에 검토해서 학습 데이터에 레이블링했기 때문에 이러한 데이터로 학습한 모델의 예측 결과 역시 임상의사결정에 충분한 증거력을 제공할 수 있다는 장점이 있다.

기존에 안저영상을 가지고 녹내장을 판독하는 접근에서는 시신경유두의 구조적 특징을 바탕으로 녹내장 여부를 판단하였기 때문에 초기 녹내장 진단에 있어서 많은 한계점을 가지고 있었다. 본 연구에서는 초기 녹내장의 주요 특징으로 나타나는 망막신경섬유층의 미세한 변화 특성까지 검출하여 녹내장 판독에 사용하는 합성곱신경망의 앙상블 방법을 적용했기 때문에 기존의 녹내장 진단 능력을 크게 향상시켰다.

높은 성능을 보이는 합성곱신경망 훈련을 위해서는 대규모의 학습 데이터를 사용해야 하는 것이 일반적 접근이다. 그러나, 본 연구에서는

상대적으로 소규모의 학습 데이터만을 가지고도 높은 성능을 보이는 연구결과를 도출한 특징이 있다. 의학 분야에서 대규모의 학습데이터를 구축하는 일은 학습 데이터의 레이블링 작업을 해당 분야의 전문가가 참여해야 하는 특성으로 인해 다른 분야보다 훨씬 더 많은 노력과 시간이 필요하다. 또한, 어떤 경우에서는 특정 질환과 관련된 실제 사례가 부족해서 의료분야의 학습 데이터 구축은 매우 어려운 일이다. 따라서 소규모의 학습 데이터로도 높은 성능을 발휘하는 방법론을 제시하는 것은 매우 의미 있는 일이다. 이러한 의료분야에 대한 데이터 구축의 제약사항을 극복하기 위해 본 연구에서는 소규모의 데이터 세트를 다변화하여 다양한 모델을 만들고, 이 모델들을 앙상블 하는 방법을 적용하여 소규모 학습 데이터의 한계를 극복하였다.

제 2 절 딥러닝 기반 녹내장 판독 시스템 활용 방안

안저영상을 바탕으로 녹내장 선별검사와 중증도 등급화를 자동화 할 수 있는 연구결과는 인공지능 기반 임상 의사 결정 지원 시스템(Clinical Decision Support System, CDSS) 개념의 S/W이다. 이러한 형태의 CDSS는 현재 널리 보급된 안저촬영기에 탑재 또는 연동하는 형태로 다양한 분야에서 활용할 수 있다.

본 연구결과를 가장 먼저 활용할 수 있는 장소는 건강검진센터이다. 현재 대부분의 건강검진 항목에 기본적으로 포함된 안저촬영 검사를 본 연구결과와 연동하여 녹내장 선별검사 자동화 서비스에 활용할 수 있을 것이다. 이렇게 활용함으로써 대규모로 이루어지는 안저촬영 결과의 판독 효율과 정확성을 높일 수 있고, 이에 따른 시간적 이득을 전문의의 2차 판독에 할애함으로써 보다 경제적이고 정확한 건강검진 결과를 얻을 수 있다.

본 연구결과는 안과 진료현장에서도 활용할 수 있다. 녹내장 질환의 만성적이고 비가역적인 특성 때문에 녹내장 환자는 주기적으로 질환의 중증도 변화를 검사해야 한다. 이 과정에서 본 연구결과를 활용하면 시간의 변화에 따른 환자의 망막 상태를 등급화한 지수를 통해 전후 비교를 쉽게 객관화할 수 있다는 장점이 있다. 따라서, 안과 진료현장에서 본 연구결과를 활용하면 다양한 효과를 기대할 수 있다.

제 3 절 기대효과

지금까지 알려진 과학 기술을 바탕으로 녹내장 질환에 대한 최적의 대응 방법은 조기에 녹내장을 발견하는 것이다. 녹내장의 확진 또는 주의 추적 관찰에 대한 의학적 의사 결정이 늦어질수록 환자가 지불해야 하는 의료비용은 크게 증가하며, 치료 효과 역시 상대적으로 많이 떨어지는 것이 일반적이다. 환자의 의료비용 증가는 국가 건강보험 지출과도 직결된 문제이기 때문에 개인적으로나 사회적인 관점에서 의료비용의 증가는 많은 부담을 가져온다. 국내 녹내장 유병률과 녹내장 관련 건강보험 의료비용 지출을 근거로 환산해보면, 녹내장 조기진단으로 최소 연간 450억 원 규모의 의료비용을 절감할 수 있을 것으로 추정한다. 따라서, 본 연구결과를 통해 많은 잠재적 녹내장 환자들에 대한 조기진단이 이루어진다면, 의료비용 절감에 상당한 파급효과를 가져올 수 있다.

녹내장을 조기에 발견하면, 환자의 실명 위험률을 95%까지 낮출 수 있으며, 조기 발견에 실패하여 실명에 이르는 경우 미국에서는 연간 35억 달러 규모의 의료비용 이외의 추가적 사회적 비용 손실을 초래한다는 보고가 있다. 본 연구결과를 통해 녹내장 환자의 조기진단이 이루어진다면 추가적 사회적 손실 비용을 절감할 수 있는 효과를 기대할 수 있다.

임상 현장에서 인공지능 기술이 겸비된 CDSS 기술을 의료진이 활용할 경우, 진단 및 판독 시간을 획기적으로 줄일 수 있다. 이렇게 줄어든 시간 자원을 환자에게 할애한다면 환자의 만족도 향상을 기대할 수 있다. 또한, CDSS를 활용한 진단을 통해 안저영상의 판독오류 역시 상대적으로 현저하게 낮출 수 있을 것으로 예상한다. 인공지능 기반 CDSS의 이러한 효과들은 결국 의료서비스 품질을 향상하고 고도화 할 수 있는 좋은 수단으로 자리 잡을 것이라 예상한다.

본 연구의 결과를 탑재한 의료기기를 제품화한다면 안저촬영 의료기기 시장에서 비교우위를 선점할 수 있고, 의료기기 매출 및 수출 등을 통한 경제적, 산업적 효과를 기대할 수 있다.

제 4 절 연구의 제한점

본 연구는 시야 검사 결과와 OCT 측정결과를 바탕으로 녹내장이 확진된 안저영상에 대해서 녹내장 등급을 레이블링하였고, 딥러닝의 학습에는 레이블링한 안저영상만을 사용하였다. 딥러닝을 위해서는 대규모의 학습 데이터를 사용해서 모델을 학습해야 하는 데 반해서 본 연구에서는 다른 연구대비 상대적으로 작은 규모의 학습 데이터로 상당히 높은 성능을 나타내는 연구결과를 제시하였다.

그러나, 이러한 노력에도 불구하고, 본 연구에는 몇 가지 제한점이 존재한다. 첫 번째로, 본 연구에 사용한 데이터는 한국인만을 대상으로 구성함에 있어서 인종별, 민족별 차이에 대한 특성을 모델에 반영하지 못했다. 전 세계인을 대상으로 본 논문에서 제시한 방법을 적용하고자 할 때 추가적인 데이터 확보와 검증이 필요하다. 두 번째로 합성곱신경

망의 입력 데이터 크기가 299x299의 해상도로 국한하여 실험을 진행하였다. 미세한 시신경 변화를 충분히 반영하기 위해서는 더 높은 해상도의 입력 영상에 대한 기계학습 실험을 고려할 필요가 있으나, 현재까지 제공하고 있는 컴퓨팅하드웨어의 한계로 인해서 입력 데이터 해상도를 확대해서 실험을 진행하는 데는 한계가 있지만, 추후 컴퓨팅하드웨어의 규격이 높아지는 것에 발맞추어 추가적인 확대 실험이 필요할 것으로 보인다. 세 번째는 녹내장의 등급을 보다 세분화하여 판독할 수 있는 합성곱신경망 모델은 의료 현장에서 임상의사결정 과정에 더 많은 정보와 도움을 줄 수 있어서 이에 대한 추가적인 실험과 연구가 필요하다.

마지막으로 본 연구에서 제시한 앙상블 방법의 최적화 방안에 관한 연구와 안저영상 전처리를 위한 추가적인 연구가 있어야 할 것이다. 또한, 안저영상에서 녹내장에서만 나타나는 영상 특징점을 강화할 수 있는 기법과 이에 대한 시계열 관점의 변화 추적 연구를 바탕으로 녹내장 예후가 보이는 정상인의 효과적인 관찰 방법에 관한 연구가 필요하다.

제 5 절 결론

본 연구를 통하여 소규모의 안저영상 데이터를 기반으로 녹내장을 판독할 수 있는 합성곱신경망 모델 구축이 가능함을 확인하였고, 이를 통해 임상 현장에서 더욱 정확하고 편리한 방법으로 초기 녹내장을 진단하는데 유용한 도구를 제공할 수 있게 되었다.

본 연구결과를 임상 현장, 특히 선별검사가 주로 이루어지는 건강검진센터 등에서 활용할 경우, 조기에 잠재적 녹내장 환자들을 선별하고 의학적 조치를 할 수 있어서, 환자의 시야 손실을 최대한 방지할 수 있

고, 이로 인한 국가 의료비용의 상대적 손실 역시 절약할 수 있다. 또한, 많은 시간과 노력이 있어야 하는 녹내장 진단에서 더욱 간편하고 정확한 방법의 도입으로 의료진의 노고를 덜어 줄 수 있으며, 이에 따른 의료서비스 품질을 향상할 수 있을 것으로 기대한다.

참 고 문 헌

1. Asaoka, R., Murata, H., Iwase, A., & Araie, M. (2016). Detecting Preperimetric Glaucoma with Standard Automated Perimetry Using a Deep Learning Classifier. Ophthalmology, 123(9), 1974–1980. https://doi.org/10.1016/j.ophtha.2016.05.029
2. Bourne, R. R. A., Taylor, H. R., Flaxman, S. R., Keeffe, J., Leasher, J., Naidoo, K., Jonas, J. B. (2016). Number of People Blind or Visually Impaired by Glaucoma Worldwide and in World Regions 1990 – 2010: A Meta-Analysis. PLOS ONE, 11(10), e0162229. https://doi.org/10.1371/journal.pone.0162229
3. Chen, X., Xu, Y., Yan, S., Wong, D. W. K., Wong, T. Y., & Liu, J. (2015). Automatic Feature Learning for Glaucoma Detection Based on Deep Learning. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 9351, pp. 669–677). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_80
4. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. arXiv preprint, 1610-02357.
5. Gupta P, Zhao D, Guallar E, et al. Prevalence of glaucoma in the United States: the 2005e2008 National Health and Nutrition Examination Survey. Invest Ophthalmol Vis Sci. 2016; 57:2577e2585.
6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90
7. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems 25 (pp. 1097–1105).
8. Lee PP, Walt JG, Doyle JJ, et al. A multicenter, retrospective pilot study of resource use and costs associated with severity of disease in glaucoma. Arch Ophthalmol. 2006;124:12e19.

9. Li, Z., He, Y., Keel, S., Meng, W., Chang, R. T., &He, M. (2018). Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. Ophthalmology.
10. Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, 1-14. Retrieved from http://arxiv.org/abs/1409.1556
11. Szegedy C, Vanhouke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. http://arxiv.org/pdf/1512.00567v3.pdf; 2015.
12. Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Rabinovich, A. (2015). Going deeper with convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1-9). IEEE. https://doi.org/10.1109/CVPR.2015.7298594
13. Szegedy, C., Ioffe, S., Vanhoucke, V., &Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI (Vol. 4, p. 12).
14. Tatham AJ, Weinreb RN, Medeiros FA. Strategies for improving early detection of glaucoma: the combined structure-function index. Clin Ophthalmol. 2014;8:611e621.
15. Tham, Y. C., Li, X., Wong, T. Y., Quigley, H. A., Aung, T., & Cheng, C. Y. (2014). Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis. Ophthalmology, 121(11), 2081-2090. https://doi.org/10.1016/j.ophtha.2014.05.013
16. Varma, R., Lee, P. P., Goldberg, I., & Kotak, S. (2011). An assessment of the health and economic burdens of glaucoma. American Journal of Ophthalmology, 152(4), 515-522. https://doi.org/10.1016/j.ajo.2011.06.004

Abstract

Deep learning model for glaucoma diagnosis and its stages classification based on fundus images

Hyeon Sung Cho
Medical Informatics, Department of Medicine
The Graduate School
Seoul National University

Introduction: This study is concerned with an ensemble method of convolutional neural networks for automatically screening tests for glaucoma and classifying the severity of glaucoma based on fundus photographs. In order to automate the glaucoma screening and classifying severity stages, we defined and trained 48 convolutional neural network models with different characteristics. Finally, the final readings were obtained through the ensemble method proposed in this study from the models in which the study has been finished and their performance was evaluated.

Methods: In this study, 4,445 fundus photographs from 2,801 patients were collected for the training of the convolutional neural network model. The collected fundus photographs were classified into a normal group (unaffected control class) and a glaucoma group by 4

ophthalmology and glaucoma specialists, and the glaucoma group was further divided into an early-stage glaucoma class and a late-stage glaucoma class by referring to the mean deviation (MD) of visual field test results. At this time, the mean deviation value of -6dB or less was classified as a late-stage glaucoma class. Also, up to one fundus photograph was used per side, left and right, for each patient. Out of the all fundus photographs, 3,460 photographs of 2,204 people were used to train the convolutional neural network model, except for the photographs with poor image quality and the ones without 100% agreement on the grade of glaucoma by 4 specialists.

The performance of the model was evaluated using the accuracy, sensitivity, specificity, and area under the receiver operating characteristic (AUROC). At this time, the performance of the proposed ensemble method in this study was compared with InceptionNet-v3 as a baseline model. The performance evaluation results of the two methods were tested using the Shapiro-Wilk normality test and the paired t-test was used to test the statistical significance of the performance differences between the two methods.

Results: The performance of the convolutional neural network ensemble method proposed in this study was evaluated separately, one related to the glaucoma screening test and one with the classification of glaucoma severity.

The accuracy of the glaucoma screening test was 96.62% (95% confidence interval [CI], 95.5 ~ 97.8%) in the ensemble method. On the other hand, the reference model using one InceptionNet-v3 model showed 93.9% (95% CI, 92.6 ~ 95.2%). The difference in performances

between the reference model and the ensemble method for glaucoma screening test accuracy was tested for statistical significance by paired t-test and the result showed that the difference of accuracy was statistically significant with the p-value of 0.000425. In terms of AUROC, the ensemble method showed 0.994 (95% CI, 0.990 ~ 0.997), and the reference model using one InceptionNet-v3 model showed 0.977 (95% CI, 0.969 ~ 0.986). The difference in performances between the reference model and the ensemble method for AUROC of glaucoma screening test was tested for statistical significance by paired t-test and the result showed that the difference of accuracy was statistically significant with p-value of 0.000966. We confirmed that the ensemble method proposed in this study has higher and more stable accuracy and AUROC for glaucoma screening compared to the reference model.

In terms of accuracy of severity classification of glaucoma, the ensemble method showed 87.7% (95% CI, 85.9 ~ 89.7%), and the reference model using one InceptionNet-v3 model showed 82.3% (95% CI, 80.2 ~ 84.1%). The difference in accuracy between the reference model and the ensemble method for glaucoma screening test was tested for statistical significance by paired t-test and the result showed that the result was statistically significant with the p-value of 0.002902. In terms of average AUROC, the ensemble method showed 0.975 (95% CI, 0.967 ~ 0.983), and the reference model using one InceptionNet-v3 model showed 0.938 (95% CI, 0.926 ~ 0.949). The difference in performance between the reference model and the ensemble method for average AUROC of glaucoma screening test was tested for statistical significance by paired t-test and the result showed that the result was statistically significant with p-value of 0.000093. We confirmed that the ensemble method proposed in this study has higher

and more stable accuracy and AUROC for glaucoma severity classification compared to the reference model.

Conclusions: The proposed ensemble method in this study using multiple convolutional neural networks, shows superior and more stable performance compared to the conventional methods in glaucoma screening test and automating severity classification based on fundus photographs. The results of this study are a clinical decision support system (CDSS) based on artificial intelligence, which can be used in various fields by installing or connecting with the currently widely used fundus camera. By using the results of this study in the fundus camera and utilizing in health check-up centers or ophthalmology clinics, it can improve the efficiency and accuracy of the reading of the fundus photograph results and focus on the second reading by the specialist with its time efficiency, ultimately obtaining more economical and accurate screening results. In addition, if the medical service utilizing the results of the present study is used more actively, the possibility of early diagnosis of potential glaucoma patients can be increased, the medical treatment expenses of the glaucoma patients can be improved, and the related medical expenses can be reduced.

Keywords: Glaucoma. Fundus Image, Deep Learning, CNN, Ensemble, AI

Student number: 2011-30639